Regressão Logística Multinomial: Um modelo à partir do comportamento longitudinal do usuário em rede social para a predição de traços depressivos

Maricy Caregnato

TESE APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
DOUTORA EM CIÊNCIAS

Programa: Ciência da Computação

Orientador: Prof. Dr. Flávio Soares Correa da Silva

Durante o desenvolvimento deste trabalho a autora recebeu auxílio financeiro da Universidade do Estado de Mato Grosso - UNEMAT

São Paulo, junho de 2018

Regressão Logística Multinomial: Um modelo à partir do comportamento longitudinal do usuário em rede social para a predição de traços depressivos

Esta é a versão original da tese elaborada pela candidata (Maricy Caregnato), tal como submetida à Comissão Julgadora.

Agradecimentos

À Deus por aceitar meses a fio minhas orações inacabadas.
À minha mãe Odalice Vera Caregnato que deve estar dançando uma milonga no céu de tanto orgulho. (in memoriam)
À minha filha Amanda Caregnato pela oportunidade de aprender todos os dias e por exercer o dom da paciência.
À minha companheira Anapaula Rodrigues Vargas por todo amor e cuidado e por ser tão tido uma síncope discricionária ao ouvir sobre meu tema.
À minha família sempre um porto seguro, pelas poucas vezes que lhes telefonei e sempre pra desabafar ou pedir algum favor.
Ao Prof. Dr. Flávio Soares Correa da Silva, pela confiança, pela oportunidade de trabalhar ao seu lado e por ser o maior incentivador na superação de meus limites e por aturar minha insistência em seguir o caminho mais longo.
À Silvana Mara Lente pela infinita disponibilidade e por todos os ensinamentos na condução do meu trabalho e por não ter desistido mesmo diante a tanta vírgula e "foi".
Aos meus grandes amigos, que sempre entenderam a minha ausência e tomaram um chopp por mim.
Ao psicometrista Luis Anunciação pelas sempre valiosas dicas e respostas a jato nas horas mais inusitadas.
Ao programador Lucas Costa pelo auxílio nos testes com o aplicativo Vivamente chegando a vibrar quando finalmente conseguimos a aprovação do Facebook.
Ao meu amigo Dionei Carrijo, pelas inúmeras horas ao telefone e vídeo conferência me auxíliando na área da psicologia.
À psicóloga e amiga Sabrina Zaffari Farias, pelo auxílio em todos os momentos do projeto.
Aos voluntários dessa pesquisa, e aos que ajudaram a divulgá-la pela sensibilidade, respeito e consideração.
Aos colegas da UNEMAT pelo auxílio prestado.

Resumo

CAREGNATO, M. **Regressão Logística Multinomial: Um modelo à partir do comportamento longitudinal do usuário em rede social para a predição de traços depressivos.** 2018. 200 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2018.

Introdução - As redes sociais tornaram-se fontes de pesquisas científicas dado ao grande volume de dados registrados pelos usuários em seus perfis sociais, possibilitando a mineração desses dados na produção de novos conhecimentos que podem contribuir na identificação de traços de comportamento depressivo. **Objetivo** - Apresentar um Modelo de Regressão Logística Multinomial, à partir do comportamento longitudinal do usuário na rede social, para a predição de probabilidades de traços depressivos. **Métodos** - De natureza aplicada, com uma abordagem quantitativa, exploratória e descritiva quanto aos objetivos, bibliográfica, de levantamento e experimental quanto aos procedimentos. Dados de postagens, curtidas e sintomas depressivos de 692 usuários da rede social Facebook, brasileiros e maiores de 18 anos foram coletados via aplicativo específico. **Resultados** - Os dados relativos aos sintomas depressivos, obtidos via Inventário de Depressão de Beck, e os dados da rede social formaram a base experimental para às análises do modelo de regressão logística multinomial, demonstrado a viabilidade na execução de um modelo capaz de predizer as probabilidade de um usuário apresentar sintomas depressivos considerando os traços depressivos e seu comportamento na rede social caracterizado por postagens e curtidas. **Conclusões** - As predições dos níveis de traços depressivos em rede social representam a capacidade das ciências atuarem de maneira interdisciplinar a contribuir para a saúde pública em prol do bem estar social.

Palavras-chave: Depressão; Rede Social; BDI-II; Mineração de Dados; Regressão Logística.

Abstract

CAREGNATO, M. **Multinomial logistic regression: a model from longitudinal user behavior in social network for depressive traits prediction.**. 2018. 200 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2018.

Introduction - Social networks have become sources of scientific research due to the large volume of data that users registered in their social profiles. This data can be mined to obtain new knowledge that may contribute to recognize depressive behavior traits. **Objective** - To present a Multinomial Logistic Regression Model, based on the longitudinal behavior of the user in social network, to predict probability of depressive traits. **Methods** - A quantitative, exploratory and descriptive approach was used for the objectives together with bibliographical, survey and experimental procedures. Data were collected through a specific app from postings, likes and depressive symptoms of 692 users from Facebook social network, all of them from over 18-year-old Brazilians the data. **Results** - The data from depressive symptoms obtained through the Beck Depression Inventory and the social network data constitute the basis for analysis of multinomial logistic regression model. This showed the feasibility of executing a model capable of predicting the probability of a user of presenting depressive symptoms considering both the depressive traits and their behavior in the social network characterized by postings and likes. **Conclusions** - The predictions of levels of depressive traits in social network represent the capacity of the sciences to act in an interdisciplinary way to contribute with public health in favor of social well-being.

Keywords: Depression; social network; Beck Depression Inventory (BDI-II); Facebook; data mining.

Sumário

Lista de Siglas e Abreviaturas

API	Interface Gráfica com o Usuário (*Application Programming Interface*)
AVC	Acidente Vascular Cerebral
BDI	Inventário de depressão de Beck (*Beck Depression Inventory*)
BFAS	Escala de vício em Facebook de Bergen (*Bergen Facebook Addiction Scale*)
CID	Classificação Internacional de Doenças
CEPH-IPUSP	Comitê de Ética em Pesquisa com Seres Humanos do Instituto de Psicologia da Universidade de São Paulo
DSM	Manual Diagnóstico e Estatístico de Transtornos Mentais (*Diagnostic and Statistical Manual of Mental Disorders*)
DM	Data Mining (*Mineração de Dados*)
EDA	Análise Exploratória de Dados (*Exploratory Data Analysis*)
IME	Instituto de Matemática e Estatística (*Institute of Mathematics and Statistics*)
GEE	Generalized Estimating Equations (*Generalized Estimating Equations*)
GLM	Modelos Lineares Generalizados (*Generalized Linear Model*)
GOF	Qualidade do Ajuste (*Goodness Of Fit*)
HADS-D	Hospital Anxiety and Depression Scale
NDDI-E	Neurological Disorders Depression Inventory for Epilepsy
HTTP	Protocolo de Transferência de Hipertexto (*Hypertext Transfer Protocol*)
KDD	Knowledge Discovery in Databases (*Descoberta do conhecimento em Bases de Dados*)
FBI	Escala de Intensidade do Facebook (*Facebook Intensity Scale*)
JSON	Notação de Objetos JavaScript (*JavaScript Object Notation*)
KDD	Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases*)
MFIS	Escala de Intensidade Multidimensional do Facebook (*Multidimensional Facebook Intensity Scale*)
MQO	Mínimos Quadrados Ordinários (*Ordinary Least Square*)
ROC	Curva Característica de Operação do Receptor (*Receiver Operating Characteristic Curve*)
SNS	Site de Rede Social (*Site de Rede Social*)

SNSs	Sites de Redes Sociais (*Sites de Redes Sociais*)
SVM	Super Vector Machine (*Site de Rede social*)
NoSQL	Armazenamento de dados não relacionais (*Not Only SQL*)
OMS	Organização Mundial da Saúde (*World Health Organization*)
PA	*Population Averaged Estimation*
REST	Transferência de Estado Representacional (*Representational State Transfer*)
RESTful	Transferência de Estado Representacional (*Representational State Transfer para serviços web*)
SQL	Linguagem de Consulta Estruturada (*Structured Query Language*)
SES	Escala de Auto-Estima de Rosenberg (*Rosenberg Self-Esteem Scale*)
SDK	Kit de Desenvolvimento de Software (*Software Development Kit*)
SVM	Máquinas de Vetores de Suporte S(*Support Vector Machine*)
SWLS	Escala de Satisfação com a Vida (*Satisfaction With Life Scale*)
TCLE	Termo de Consentimento Livre e Esclarecido (*Consent Form Free and Informed*)
TDM	Transtorno Depressivo Maior (*Major Depressive Disorder*)
TI	Tecnologia da Informação (*Information Technology*)
TIC	Tecnologia da Informação (*Information and Communication Technology*)
UGC	Conteúdo gerado pelo usuário (*User Generated Content*)
URL	*Uniform Resource Locator*
VIF	*Variance Inflation Factor*
URI	*Uniform Resource Identificator*
UNEMAT	Universidade do Estado de Mato Grosso
UFMT	Universidade Federal de Mato Grosso

Lista de Símbolos

μ	Média
η	Função de ligação canônica
α	Constante
β	Coeficiente de cada variável explicativa
$\hat{Y}$	Valor esperado da variável
p	Probabilidade de ocorrência do evento de interesse
Z	*Logit*
H_0	Hipótese nula
H_1	Hipótese alternativa
$s.e$	Erro padrão *standard error*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O interesse em apresentar um modelo de regressão logística capaz de contribuir para a predição de traços depressivos exige uma prévia contextualização quanto as taxas de incidência e prevalência desta patologia no mundo e no Brasil, as quais demonstram a emergente necessidade de estratégias e ações para a sua redução e controle. Acrescido às considerações premilinares sobre depressão este capítulo abrange os objetivos, as contribuições e a organização desta pesquisa.

1.1 Considerações Preliminares

De acordo com a Organização Mundial da Saúde OMS [2017] os dados mundiais destacam que a depressão é responsável por 7,5% da incapacidade humana e a principal causa de mortes por suicídio, com cerca de 800 mil casos por ano. Com a estimativa para 2020 de se tornar a segunda causa de mortes no mundo, ficando somente atrás das doenças cardíacas.

Segundo Baldwin e Birtwistle [2002] a combinação de fatores genéticos, ambientais e psicológicos podem ser a causa do desenvolvimento da depressão. E, esta quando não diagnosticada e tratada tende a ser crônica e associada a um maior grau de incapacidade em realizar atividades diárias, conforme aponta a OMS [2017].

Pessoas deprimidas têm frequentemente pensamentos mórbidos e a taxa de suicídio entre depressivos é 30 vezes maior do que a média da população em geral. O que leva a depressão ser considerada em várias partes do mundo como uma das doenças com a mais alta taxa de mortalidade segundo Guedes [2015].

Conforme os resultados da pesquisa realizada por Ferrari *et al.* [2013] a depressão é mais freqüente na América do Sul e no sul da Ásia do que na Europa Ocidental, com maior incidência entre as mulheres.

Já registros da OMS [2017] demonstram que o Brasil tem a maior taxa de pessoas com depressão da América Latina, com uma média que supera os índices mundiais, com registro de 322 milhões de pessoas no mundo acometidas por depressão, 18% a mais do que há dez anos; o número representa 4,4% da população mundial e 5,8% dos brasileiros.

O período de pico de desenvolvimento segundo Hallowell e Ratey [2005] é entre 25 a 44 anos, o que leva a estimar que 10% a 25% das mulheres e 5% a 12% dos homens irão sofrer de algum episódio depressivo ao longo da vida. Já demonstrado nos Estados Unidos e na Europa as maiores diferenças entre homens e mulheres diagnosticados.

Essa doença pode afetar de forma negativa as relações familiares da pessoa, o emprego ou a vida escolar, o sono, as refeições e a saúde em geral, entre 2 a 7% dos adultos com depressão morrem de suicídio, conforme explicita Richards *et al.* [2014].

Frente aos elevados índices de pessoas acometidas com depressão, a ciência não pode se furtar na busca de elementos que possam contribuir para o controle e a redução dos mesmos, possibilitando assim o alcance de uma condição saudável à população brasileira.

Neste contexto é que se ocupa da ciência da computação associada a ciência matemática e estatística para demonstrar que é possível de maneira interdisciplinar estas áreas de conhecimento

contribuir para um prévio diagnóstico a partir da identificação de traços depressivos em redes sociais.

Vale ressaltar aqui, que dado a esta marcante expressão comportamental dos usuários nos Sites de Redes Sociais (SNSs), estes se tornaram grandes fontes de dados que podem ser transforma- dos em informações relevantes em diversas áreas, inclusive para novas produções científicas. A exemplo do trabalho de Sheu *et al.* [2015] os quais se utilizaram de textos das postagens e curtidas na rede social Facebook com o objetivo de inferir traços de personalidade do usuário por meio de estilos de escrita e número de curtidas.

Destaca-se ainda o estudo de Kim e Yang [2017] onde por meio de dados extraídos de likes, comentários e compartilhamentos aferiu comportamentos distintos de usuários no Facebook. Sendo nesta mesma linha desenvolvido o estudo de Jamil [2017] onde dados de postagens públicas do Twitter foram utilizados com o intuito identificar usuários que sofram de depressão ou estejam em risco de depressão, usando técnicas de mineração de texto.

Por outro lado, área da regressão logística é bastante utilizada em estudos aplicados à área da saúde, tanto a regressão linear simples e múltipla, quanto a logística binária e multinomial, seja em painel ou multinível. Essencialmente na predição de algum fator que poderá ajudar a entender como as diversas patologias se manifestam por meio do comportamento, nos mais diversos tipos de doenças, em destaque o câncer, as doenças cardíacas e os distúrbios mentais.

Isto porque, o objetivo da regressão logística é gerar uma função matemática cuja resposta permita prever a probabilidade de uma observação pertencer a categoria mínimo, leve, moderado ou grave de traços depressivos representados pela variável *label4*, em razão do comportamento do usuário na rede social por meio de postagens ou curtidas, demais atributos pessoais e sintomas depressivos.

Nesta linha de raciocínio, apropriando-se da regressão logística e de dados de rede social esta pesquisa se justifica frente sua contribuição no campo científico com a valorização das expressões de sentimentos e comportamentos do ser humano em busca de novas ciências. Neste contexto, este estudo se delineia, articulando ciências em prol da humanidade.

1.2 Objetivos

1.2.1 Objetivo Geral

Apresentar um Modelo de Regressão Logística Multinomial, à partir do comportamento longitudinal do usuário na rede social, para a predição de probabilidades de traços depressivos.

1.2.2 Objetivos Específicos

Coletar dados na rede social a partir do aplicativo Vivamente e Inventário de Depressão de Beck para a extração de um conjunto de informações legalizadas e adequadas ao modelo de regressão proposto;

Preparar os dados coletados a partir da mineração dos dados com a classificação da severidade depressiva dos usuários com a utilização da Escala de Depressão de Beck BDI-II;

Transformar os dados coletados em forma longitudinal por meio de um Script MongoDB possibilitando extrair a polaridade dos sentimentos;

Experimentar o modelo proposto a partir de Test T, Coeficiente de correlação, Razão de Chance (RC) e Teste de Wald para o alcance da predição de probabilidades de traços depressivos.

1.3 Contribuições

As principais contribuições deste trabalho são as seguintes:

- Registra-se que as redes sociais se tornaram fontes inesgotáveis de pesquisas científicas dado ao grande volume de dados registrados pelos usuários em seus perfis sociais, o que permite

à ciência produzir novos conhecimentos a partir de estudos que agreguem estas informações como fonte de pesquisa;

- Este estudo poderá contribuir diretamente para um despertar de novos interesses da área da Ciência da Saúde para a associação de informações de respectivos pacientes como fonte diagnóstica e prognóstica no tratamento da depressão;
- Também contribuiu para demonstrar que por meio do modelo aplicado é possível proceder a mineração de dados substanciais e elementares que indicarão traços depressivos em usuários de redes sociais, o que pode ser associado à saúde no tratamento do quadro depressivo de determinado paciente;
- Além, de contribuir para em nível de políticas públicas de saúde para redução do índice de pessoas depressivas a partir de estratégias educativas por meio das redes sociais. Ou ainda, no cruzamento de fontes de informações, entre os Sistemas de Informações do Centro de Atenção Psicossocial e as informações constantes nos perfis dos usuários das redes sociais;
- Ao identificar o padrão de mudança de comportamento longitudinal de cada grupo de usuários pertencentes às diferentes categorias de sintomas depressivos, sendo eles mínimo, leve, moderado e grave contribui para a classificação da severidade da doença em pacientes depressivos, podendo contribuir para o tratamento adequado ao quadro diagnóstico;
- As predições dos níveis de traços depressivos por meio dos atributos referentes ao comportamento do usuário na rede social representam a capacidade das ciências atuarem de maneira interdisciplinar em prol do bem estar social;
- Os ajustes dos modelos de dados representam uma variação das contribuições técnicas onde o pesquisador consegue se apropriar de métodos e técnicas existentes tornando-os exequíveis em outras realidades.

1.4 Organização do Trabalho

Este trabalho está organizado em capítulos como se apresenta:

O Capítulo 1 dispõe sobre esta introdução com as considerações preliminares, os objetivos, as contribuições e a organização desta disposição gráfica desta pesquisa.

No Capítulo 2 corresponde ao Marco Teórico o qual abrange a teorização da depressão com destaque a sua conceituação, seus sintomas e diagnóstico, com a finalidade de despertar no leitor quão emergente são ações que possibilitem o controle e o tratamento da mesma.

Ainda no campo da teorização tem-se sobre a regressão e seus modelos, com ênfase no Modelo de Regressão Logística Multinomial, aplicado nesta pesquisa; bem como, destaque para alguns estudos desenvolvidos no contexto brasileiro.

Para contribuir com o embasamento teórico ocupou-se em fundamentos teóricos sobre a Mineração de Dados e trazer registros científicos sobre a Tecnologia da Informação e Comunicação à serviço da saúde, destacando as redes sociais como fontes de informações relevantes para coletar dados relacionados a traços de comportamento depressivo. Aqui ainda se apresenta vários estudos desenvolvidos no Brasil que se assemelham a este estudo.

O Capítulo 3 corresponde aos Procedimentos Metodológicos do estudo, demonstrando o tipo de pesquisa aqui aplicada correspondente a uma pesquisa experimental, os métodos e técnicas aplicados para a coleta de dados e para as análise dos resultados alcançados.

E por fim o Capítulo 4 apresenta os resultados e discussões da pesquisa com ênfase na Análise e resultados do Modelo de Regressão proposto. Seguido das considerações finais, referências, apêndices e anexos.

Capítulo 2

Marco Teórico

Esse capítulo aborda primeiramente, de forma sucinta, os caminhos percorridos pelos termos melancolia e depressão ao longo da história, sem a pretensão de efetuar um estudo completo mas sim um breve resumo da descrição histórica com o propósito de complementar o entendimento do tema. Em seguida foi construída uma base teórica sobre Tecnologia da Informação e Comunicação a serviço da saúde com destaque à Mineração de Dados, e por fim sobre Regressão Logística e suas variações. Trazendo a cada subtítulo estudos que demonstram a aplicabilidade teórica na prática, testados a cada pesquisa desenvolvida no contexto mundial e nacional.

2.1 Depressão

Neste subtítulo trata-se sobre a depressão, onde em linhas gerais a conceitua à partir de sua classificação e definição ao longo do tempo. Seguindo com a classificação dos sintomas depressivos adjacentes e a exposição da escala selecionada para mensurar os sintomas depressivos a ela inerentes; bem como, estudos realizados sobre depressão a partir das tecnologias.

2.1.1 Conceito de depressão

Beck e Alford [2009] acentuam que já decorreu mais de 2 mil anos desde que a depressão foi reconhecida como um transtorno e até hoje não foi encontrada uma explicação plenamente satisfatória de suas características intrigantes e paradoxais; ainda existem importantes questões não resolvidas sobre a sua natureza, classificação e etiologia. A concepção etiológica dos quadros clínicos demarcam um problema e diferentes teorias passaram a disputar o domínio desse território, segundo Verztman [1995].

Registros científicos dão conta que no início do Século XIX, Pinel conceituou a melancolia de uma forma muito semelhante à atual, sendo semelhantes às anotações autobiográficas contemporâneas como as de Clifford W. Beers, Beers [2010]. Porém, Verztman [1995] afirma que a depressão enquanto transtorno mental de fundamentou apenas na Década de 50 do Século XIX, e foi responsável por alavancar um novo conceito do que até então vinha sendo conceituado como melancolia.

O termo depressão costuma ser empregado nas ciências da saúde a partir de três (03) diferentes conceitos: o primeiro a define como um sintoma (estado de tristeza ou humor deprimido); o segundo como um transtorno, em que os sinais e sintomas envolvidos no transtorno depressivo maior decorrem de uma condição médica geral ou do uso de alguma substância; e o terceiro como uma psicopatologia, denominada Transtorno Depressivo pela OMS [2017] e pelo CID-10, de acordo com Caetano [1993] e Transtorno Depressivo Maior pelo DSM-IV [1994], descrito em Powell *et al.* [2008].

Assim, em linhas gerais a depressão tanto na Classificação Internacional das Doenças (CID-10) quanto no Manual Diagnóstico e Estatístico de Transtornos Mentais (DSM-IV) é conceituada como um transtorno mental, caracterizado por alguns sintomas em comum e inerentes a ela, conforme Gruenberg *et al.* [2005]. Já a OMS [2017] a define como "[...] doença caracterizada por tristeza

persistente e perda de interesse em atividades normalmente apreciadas, acompanhada por uma incapacidade de realizar atividades diárias, persistindo por pelo menos duas semanas".

Cabe destacar, que nessa pesquisa, o termo "depressão"foi utilizado para referir-se ao conjunto de sintomas que caracterizam o Transtorno Depressivo Maior (TDM), como exposto anteriormente e apresentados em detalhes a seguir.

2.1.2 Sintomas da depressão

Dentre as classificações dos sintomas depressivos, uma divisão bem aceita refere-se ao modo cognitivo, o fisiológico e o comportamental. O cognitivo envolve sentimentos como o humor deprimido, anedonia, dificuldade de concentração e tomada de decisões e ideação suicida. O fisiológico compreende a fadiga, a alteração de sono e apetite, redução do interesse sexual e agitação. O comportamental envolve isolamento social, choro, comportamentos suicidas, lentificação ou agitação psicomotora, segundo Cunha [2001].

Os sintomas utilizados hoje nos diagnósticos de depressão, conforme Beck e Alford [2009] se encontram nas descrições antigas tais como: humor perturbado, autopunição, comportamento autodepreciativo, desejo de morrer, sintomas físicos e vegetativos e delírios de ter cometido pecados imperdoáveis. As descrições da depressão mencionadas incluem as características típicas da condição do transtorno de depressão, poucos são os transtornos psiquiátricos que possuem descrições clínicas tão constantes ao longo de sucessivas épocas da história.

Alguns desses sintomas citados, sem uma classificação específica, seguem versados em consonância com o descrito pelo BDI-II:

O sintoma do humor deprimido é caracterizado por sentimento de tristeza e mal estar generalizado, conforme Powell *et al.* [2008]. Onde muitas vezes, segundo Beck e Alford [2011] o interesse ou prazer em atividades cotidianas diminui, de maneira que esses sentimentos estejam presentes diariamente por duas semanas ou mais para cumprir os critérios do DSM-V para caracterizar o TDM.

Considera-se conforme explica Beck e Alford [2011] que comumente a pessoa expressa o sentimento em termos somáticos como "tenho um nó na garganta"ou "tenho uma sensação de vazio no estômago"ou "tenho uma sensação de peso no peito", ou, através de adjetivos como triste, infeliz, solitário ou entediado. Ainda em Powell *et al.* [2008], a redução na capacidade de sentir prazer está associada a falas como: "já não tenho mais prazer naquilo que gostava", "perdi o interesse pelas coisas do mundo e pelas pessoas"ou "o mundo parece cinza".

Vale considerar que a maioria dos pacientes deprimidos relatam algum grau de tristeza ou infelicidade. Alguns apresentam períodos flutuantes em que se sentem tristes, enquanto outros ficam incapacitados pela intensidade do sentimento de infelicidade, conforme Beck e Rush [1979].

A perda de interesse é um processo bem comum entre as pessoas depressivas e muitos a consideram a característica central de sua doença, e geralmente relatam ao menos uma perda parcial da satisfação, podendo iniciar por algumas atividades e à medida que a depressão evolui se propaga para praticamente todas as atividades executadas pela pessoa, conforme Beck e Alford [2011].

Outro sintoma depressivo, consiste na alteração do apetite ou peso, onde a pessoa depressiva pode ter uma acentuada perda ou ganho de peso (como 5 % do seu peso corporal em um mês) ou uma mudança no apetite; comer pouco ou comer muito. No primeiro caso, algumas pessoas nunca sentem fome, podem ficar longos períodos sem comer, podem esquecer de comer ou se comem uma pequena quantidade de alimento pode ser suficiente. Já no segundo caso, algumas pessoas tendem para um aumento do apetite e podem ganhar uma quantidade significativa de peso, elas podem preferir certos tipos de alimentos, como doces ou carboidratos, conforme Lee *et al.* [2000].

A perda do apetite e do interesse sexual são frequentes nos primeiros sintomas da depressão, ambos parecem ser manifestações da perda generalizada de prazer do paciente em quaisquer atividades. Porém, à medida que a depressão passa, o apetite retorna, conforme Beck e Alford [2011].

Vale relatar que alguns pacientes comem demais e engordam ao ficarem deprimidos, enquanto outros engordam enquanto moderadamente deprimidos e emagrecem quando em depressão aguda. Beck e Rush [1979] registram que os pacientes frequentemente se mostram apreensivos a respeito

de ganharem peso, visto que fazer regime para perder peso é um processo difícil para um paciente deprimido.

A Alteração no sono, que também se configura como um sintoma depressivo é um dos mais marcantes nesta patologia. A maioria das pessoas deprimidas apresentam algum tipo de distúrbio do sono; esses problemas incluem dificuldade de adormecer, sono agitado e o despertar cedo demais pela manhã. Em geral, a pessoa recupera seu padrão de sono normal depois que a depressão se reduz, conforme Beck e Rush [1979].

Quase diariamente, a pessoa pode dormir excessivamente (hipersonia) ou não dormir o suficiente (insônia) sendo o último mais recorrente em pessoas deprimidas. Os sintomas de insônia incluem dificuldade em adormecer, dificuldade para ficar dormindo e/ou acordar muito cedo de manhã. Já a hipersonia é um tipo menos comum de distúrbio do sono que pode incluir o sono por períodos prolongados durante a noite ou aumento durante o dia, o sono pode não ser repousante e a pessoa pode se sentir lenta apesar de muitas horas de sono, segundo Roehrs *et al.* [1994].

Embora os pacientes deprimidos durmam menos que as pessoas que não apresentam depressão, muitos exageram a extensão de sua insônia. O paciente que declara haver passado toda a noite em claro terá tido, possivelmente, um sono leve, boa parte do tempo. A minimização feita pelo paciente do tempo real de sono vem geralmente acompanhada da crença de que precisa de maior número de horas de sono do que são efetivamente necessárias, conforme Beck e Alford [2011].

O sentimento de culpa é reconhecido como um sintoma da depressão, e, algumas pessoas podem sentir-se culpadas por seus sentimentos ou desejos mais do que por ações especiais. Esse sentimento de culpa relaciona-se com frequência ao mecanismo da pessoa em assumir uma parcela irreal da responsabilidade pelo comportamento de terceiros, conforme Beck e Rush [1979].

Pessoas deprimidas podem ter sentimento de culpa que vão de um nível normal até delírios. As pessoas deprimidas julgam a si mesmas de maneira muito negativa, não realista, manifestando preocupação com fracassos passados, singularizando eventos triviais ou acreditando que pequenos erros possam confirmar sua incompetência. Podendo ter um senso realista de responsabilidade pessoal e ver as coisas além de seu controle como sendo culpa delas, logo, a autodepreciação é também algo comum na depressão e pode levar a um declínio quando combinado com outros sintomas, segundo Beck e Alford [2009].

Em relação ao sintoma "alteração na atividade psicomotora" as pessoas próximas podem perceber que o nível de atividade da pessoa não está normal. Pode ser excessivamente ativa (agitação psicomotora) e andar pela sala, esfregar as mãos ou brincar com roupas e objetos, ou ser muito lenta (retardo psicomotor) e se mover lentamente, desviar os olhos, permanecer em uma cadeira e falar devagar, falando pouco, dizendo que seus braços e pernas estão pesados, conforme Beck e Alford [2011].

O conteúdo dos pensamentos das pessoas com retardo psicomotor parece girar em torno da aceitação passiva do seu destino. Já a pessoa agitada tem dificuldade em aceitar ou suportar a tortura prevista, parecendo representar tentativas desesperadas de lutar contra o mal que se aproxima, conforme Beck e Alford [2011].

A principal característica das pessoas agitadas é a atividade incessante, transmitindo uma sensação de inquietação e perturbação. A noite saem da cama e andam de um lado para outro sem parar. A agitação é também manifestada por queixa, gemidos e lamúrias, Beck e Alford [2011]. Relatos pessoais de sensação de inquietação ou lentidão não contam para os critérios de diagnóstico, segundo Beck e Alford [2009].

No sintoma relacionado à Fadiga ou perda de energia a pessoa que mostra comportamentos depressivos geralmente reclama de cansaço extremo depois de um esforço que para outras pessoas seria mínimo, essa queixa pode ser interpretada por familiares e amigos como avolia , conforme Powell *et al.* [2008].

As pessoas podem apresentar grande dificuldade em realizar tarefas vitais, tais como comer, ir ao banheiro ou tomar remédios, e, apesar de estarem aptas, não sentem nenhum estímulo para agirem, conforme Beck e Alford [2011].

No cotidiano, a pessoa terá extrema fadiga, cansaço ou perda de energia. Vale destacar que uma

pessoa pode se sentir cansada sem ter feito qualquer atividade física e tarefas do dia a dia tornam-se cada vez mais difíceis; as tarefas de trabalho ou tarefas domésticas se tornam muito cansativos levando o paciente a interpretar que seu trabalho lhe causa sofrimento. A pessoa pode ficar indecisa ou têm dificuldade para pensar ou se concentrar de acordo com Clark e Beck [2009].

Logo, problemas com a memória e distração também são comuns, estas questões causam dificuldades significativas às pessoas envolvidas em atividades intelectualmente exigentes, como os estudos ou o trabalho, especialmente em áreas complexas, segundo Beck e Alford [2009].

Por fim, trata-se quanto ao sintoma relacionado ao Pensamento de morte e suicídio, onde os desejos suicidas têm historicamente sido associados ao estado depressivo, embora também ocorram em indivíduos não depressivos. Ressalta-se que as ideações suicidas ocorrem com muito mais frequência nos pacientes deprimidos, conforme Beck e Alford [2011].

A pessoa pode ter recorrentes pensamentos sobre a morte (que não seja o medo de morrer) ou suicídio (com ou sem um plano) ou pode ter feito uma tentativa de suicídio. A frequência e intensidade dos pensamentos sobre suicídio podem variar desde acreditar que os amigos e familiares se sentiriam melhores se ela estivesse morta ou pensamentos frequentes em cometer suicídio (geralmente relacionados ao desejo de cessar a dor emocional), até planos detalhados sobre como o suicídio seria realizado. Sendo que àqueles com pensamentos suicidas mais severos podem até fazer planos específicos e decidir um dia e local para a tentativa de suicídio, conforme Beck e Alford [2011].

Como pode ser observado, na maioria dos sintomas não há nada de característico em cada uma das sensações descritas que determine por si só provas da existência de um transtorno mental. Sofrimento ou humor deprimido, variações no peso, apetite ou sono na atividade psicomotora são possíveis a todos dentro dos limites de normalidade e não denotam um transtorno depressivo em casos isolados. Porém, a frequência, intensidade e variação separam o que seria uma mente saudável de uma possível experiência depressiva, de acordo com Clark e Beck [2009].

Em relação à gravidade, pode ser classificada, de modo genérico, em tipo leve, moderado ou grave, com características psicóticas, em remissão parcial ou em remissão completa; segundo o DSMV. No TDM leve o sofrimento é manejável e os sintomas geram pouco prejuízo no funcionamento social ou profissional, o TDM grave se justifica por um número de sintomas muito maior que o requerido para estabelecer o diagnóstico, o sofrimento é grave, não manejável e há um importante prejuízo na funcionalidade, o tipo moderado situa-se entre os dois mencionados, já o TDM com características psicóticas se define pela presença de delírios e/ou alucinações, conforme Nogueira *et al.* [2014].

Como já discorrido, a depressão é um dos transtornos mentais mais prevalentes em vários países, segundo a OMS [2017]. E, acompanhando os registros científicos sobre esta patologia, passa-se a demonstrar algumas formas de diagnóstico da depressão, em específico algumas escalas de medidas na detecção de classificação depressiva.

2.1.3 Diagnóstico da depressão à partir de escalas de classificação: Inventário de Depressão de Beck (BDI)

O diagnóstico da depressão tem como referência a descrição das experiências por parte da pessoa, por meio de entrevista, e a posterior avaliação do estado mental, segundo descreve Baldwin e Birtwistle [2002]. Uma avaliação diagnóstica pode ser realizada por um clínico geral devidamente treinado, ou por um psiquiatra ou psicólogo, e inclui uma infinidade de fatores que podem ser levados em consideração para um diagnóstico mais preciso, conforme Patton [2015].

O exame de saúde mental pode incluir o uso de uma escala de classificação, como a *Hamilton Rating Scale for Depression*, conforme expõe Zimmerman *et al.* [2004] ou o *Suicide Behaviors Questionnaire-Revised*, segundo Osman *et al.* [2001]. Outras escalas como a Escala de Avaliação de Depressão de *Montgomery-Asberg*, Escala de Avaliação de Melancolia de *Bech-Rafaelsen*, Escala de Auto avaliação de Depressão de *Zung*, Inventário de Auto-avaliação de *Wakedield*, a Escala de Avaliação de Depressão de *Carrol* e Inventário de Depressão de *Beck*, também podem ser utilizadas na investigação de sintomas depressivos, conforme exibem Calil e Pires [1998].

A pontuação em uma escala de classificação por si só não é suficiente para diagnosticar depressão para satisfação da regulamentação do DSM ou CID, mas fornece uma indicação da gravidade dos sintomas por um período de tempo, de modo que uma pessoa que pontue acima de um determinado ponto de corte pode ser mais detalhadamente avaliado para um diagnóstico de transtorno depressivo, conforme Sharp e Lipsky [1998].

Para ser útil, uma escala investigativa de depressão deve ser capaz de explicar os principais sintomas da depressão em termos de princípios comportamentais estabelecidos empiricamente. O DSM-IV, lista os principais sintomas da depressão, o número e a duração dos sintomas que devem estar presentes para garantir um diagnóstico específico, que incluem feições deprimidas ou tristes, redução do interesse ou prazer em realizar atividades (anedonia), alterações de apetite (tanto ganho como perda de peso), alterações de sono (tanto insônia como excesso de sono), redução geral do nível de atividades (retardo psicomotor), agitação ou ansiedade, fadiga ou perda de energia, sentimentos de inferioridade e/ou culpa contínua acompanhados por autocrítica, recordação seletiva ou atenção para eventos negativos, distorção cognitiva e ideação suicida, conforme Dougher e Hackbert [2003].

Aliando os princípios estabelecidos no DSM destaca-se o Inventário de Depressão de Beck (BDI), compreendo um escala avaliada e validada no Brasil, a qual foi utilizada nessa pesquisa em sua variação BDI-II, criada por Beck. Nesta seara, cabe delinear que *Aaron Temkin Beck* é um psiquiatra norte-americano e professor emérito do departamento de psiquiatria na Universidade da Pensilvânia, ele é conhecido como pai da Terapia Cognitiva e inventor das Escalas de Beck,[1] que são vastamente utilizadas.

Beck é famoso por sua pesquisa em psicoterapia, psicopatologia, suicídio e psicometria, que levou à criação da Terapia Cognitiva, pelo qual recebeu o Prêmio Lasker,[2] também é o criador do BDI, um dos instrumentos mais utilizados como métrica dos sintomas da depressão. Beck acreditava que a depressão manifestava-se por causa das visões negativas não realistas sobre o mundo, onde as pessoas deprimidas teriam uma percepção negativa em três áreas, que são estabelecidas como a tríade depressiva, desenvolvem visões negativas sobre: elas mesmas (o *self*), o mundo e seu futuro, tendo grande relevância no desenvolvimento do quadro depressivo, um exemplo da tríade, retirado de Brown *et al.* [1995] é o caso do estudante que obteve maus resultados nas provas:

- O estudante apresenta pensamentos negativos sobre o *mundo*, assim ele passa a acreditar que não gosta das aulas.
- O estudante apresenta pensamentos negativos sobre seu *futuro*, pois pensa que não será aprovado na disciplina.
- O estudante apresenta pensamentos negativos sobre o seu *self*, já que acredita que não merece estar na faculdade.

O desenvolvimento do BDI reflete estes problemas na sua estrutura com frases como: "eu perdi todo o interesse em outras pessoas", para refletir o mundo; "eu me sinto desencorajado sobre o futuro", para refletir o futuro; e "eu me culpo por tudo de ruim que acontece", para refletir o *self*. A visão da doença depressiva como mantida por cognições negativas intrusivas tem particular aplicação na Terapia Cognitivo Comportamental, que surgiu com o objetivo de corrigir os pensamentos distorcidos e aliviar os sintomas depressivos, segundo Beck *et al.* [1996].

O BDI é uma escala de autoavaliação que teve a sua validade extensamente estudada e documentada na literatura especializada, o instrumento original, o *Beck Depression Inventory*, foi desenvolvido em 1961 por *Beck* e colaboradores para avaliar a sintomatologia depressiva, apoiado no paradigma teórico de que as cognições depressivas seriam as alterações psicopatológicas mais importantes para mensurar o quadro clínico de depressão, de acordo com Clak e Beck [1999].

A fácil aplicação do BDI e sua positiva aceitabilidade pelos usuários ampliou ainda mais o seu uso em um grande número de estudos, o BDI recebeu várias revisões do próprio autor para

[1] Beck Scales for Adults and Children. Beck Institute for Cognitive Threapy and Research.
[2] The Lasker Awards.

aperfeiçoar as necessidades clínicas e de pesquisa. Em 1974, foi desenvolvida uma forma abreviada com 13 itens; em 1978, a escala de 21 itens foi revisada, diferindo da original quanto ao tempo de referência da avaliação, "última semana"em vez de "hoje", e por pequenas alterações na redação de seus itens, conforme Clak e Beck [1999].

O BDI original foi apoiado nas afirmações descritivas típicas dos sintomas que eram frequentemente relatados por pacientes psiquiátricos com depressão e, somente às vezes, por pacientes psiquiátricos sem depressão, as observações clínicas e as descrições dos pacientes foram sistematicamente consolidadas em 21 itens representativos de sintomas e atitudes depressivos, esses itens foram organizados de acordo com a intensidade do conteúdo das diferentes afirmações, a cada item foi atribuído um valor de acordo com uma escala de 4 pontos, variando de 0 a 3 em intensidade.

A versão original foi construída para ser aplicada por entrevistadores treinados que liam em voz alta as afirmações aos pacientes e em seguida, os pacientes selecionavam as afirmações de cada item que melhor se ajustavam ao seu estado atual de humor. Os entrevistadores geralmente levavam de 10 a 15 minutos para aplicar o instrumento e obter a pontuação total mediante a soma dos escores atribuídos pelos pacientes em cada um dos 21 itens, segundo Beck *et al.* [1996].

Desde 1971, *Beck* e seus associados começaram a empregar uma versão modificada do BDI, onde este eliminava expressões alternativas para os mesmos sintomas, como afirma Beck e Rush [1979].

O BDI-IA substituiu o instrumento original (BDI), e a edição de 1993 desse manual incluiu pequenas revisões da faixa de pontuação recomendadas para interpretar o nível de intensidade dos sintomas depressivos, de acordo com Beck *et al.* [1996].

Os sucessivos lançamentos do DSM-III-R e do DSM-IV, destacaram a necessidade de uma nova medida psicológica de depressão que avaliasse os sintomas conforme os critérios desses sistemas de classificação. Além disso, alguns dos sintomas originais do BDI que foram tipicamente observados, principalmente a mudança da auto imagem, perda de peso e preocupações somáticas, tornaram-se gradualmente menos úteis com o passar dos anos para a avaliação da intensidade da depressão, sendo assim iniciou-se a criação de uma nova versão do DBI, denominada DBI-II.

Beck e sua equipe começaram o trabalho piloto em 1994 com o Inventário de Depressão de Beck segunda edição (BDI-II), sendo esse um instrumento de auto aplicação composto por 21 itens, cujo objetivo é medir a intensidade da depressão em adultos e adolescentes a partir dos 13 anos de idade. O questionário completo e detalhado com os 21 itens encontra-se no G. Essa versão do BDI-II foi desenvolvida para avaliar os sintomas correspondentes aos critérios diagnósticos dos transtornos depressivos descritos no DSM-IV da APA.

Registra-se que após 35 anos de experiência e pesquisa com o BDI foi considerado propício revisar e modernizar a versão modificada do BDI-IA, descrito em Gorenstein *et al.* [2011].Assim, o BDI-II é uma ferramenta de avaliação consolidada no meio científico internacional, para medir a presença e gravidade de sintomas depressivos, tanto da população clínica como da população geral. Ao longo de sua existência, esse inventário alcançou grande popularidade entre pesquisadores e clínicos, sendo utilizado em numerosos estudos científicos sobre depressão no mundo todo.

No Brasil o BDI-II foi traduzido e adaptado por Cunha [2001], e passou por diversos estudos de validação. Sua fácil aplicação e aceitabilidade pelos usuários ampliou ainda mais o seu uso em diversas pesquisas.

Embora a reformulação do BDI-II tenha claros objetivos para se adequar aos critérios diagnósticos do DSM-IV, este instrumento não serve para fazer diagnóstico psiquiátrico por não envolver avaliação clínica, entretanto, ele pode ser valioso para documentar a presença de sintomas depressivos e avaliar a sua gravidade, tanto em pessoas da população geral quanto em pacientes deprimidos diagnosticados clinicamente, conforme afirma Gorenstein *et al.* [2011].

Nesta versão do BDI-II, os quatro itens: perda de peso, mudança na autoimagem, preocupações somáticas e dificuldade de trabalhar foram retirados e substituídos por outros 4 novos itens: agitação, desvalorização, dificuldade de concentração e falta de energia, com o objetivo de identificar sintomas típicos de depressão grave, foram modificados dois itens para incluir tanto o aumento quanto a diminuição de apetite e sono, também foram reescritas algumas afirmações (ou alternati-

vas) utilizadas para avaliar outros sintomas.

2.1.4 Abordagem bifatorial da depressão e o Inventário de Depressão de Beck (BDI-II)

A depressão pode ser entendida como tendo dois componentes: o componente afetivo (humor) e o componente físico (ou somático) por exemplo, perda de apetite. A BDI-II reflete esta abordagem e pode ser dividida em duas subescalas. O propósito destas subescalas é determinar a causa primária da depressão, segundo Beck *et al.* [1996].

A subescala afetiva contém 8 itens: pessimismo, perdas passadas, sentimentos de culpa, sentimentos de punição, auto-desprezo, autocrítica, pensamentos ou desejos suicidas e pensamentos de desvalor. A subescala somática consiste de outros 13 itens: tristeza, alterações no apetite, perda de prazer, choro, agitação, perda de interesse, cansaço ou fadiga, indecisão, perda de energia, alterações nos padrões de sono, irritabilidade, dificuldades de concentração e diminuição da libido. As duas subescalas são moderadamente correlacionadas a 0,57, o que sugere que os aspectos físicos e psicológicos da depressão são relacionados ao invés de completamente distintos Storch *et al.* [2004].

Embora Conforme Beck *et al.* [1996], o desenvolvimento da BDI tenha sido um importante acontecimento em psiquiatria e psicologia, bem como, representado uma mudança na visão da depressão, pois transitou de uma abordagem freudiana e psicodinâmica para uma abordagem guiada pelos pensamentos ou cognições. Assim como outros inventários ou escalas de auto-relato a BDI pode possuir um certo viés, nos quais os resultados podem ser facilmente exacerbados ou reduzidos pelo usuário que as responde.

Da mesma maneira que os demais questionários, a forma na qual o instrumento é administrado pode causar um efeito no resultado final. Se um usuário completar o questionário da escala na presença de outras pessoas, em um ambiente clínico, por exemplo, as expectativas sociais podem criar resultados divergentes quando comparados à aplicação através de envio.

Em usuários com manifestações físicas associadas, o peso do BDI em traços físicos como fadiga, por exemplo, pode artificialmente incrementar os resultados devido a sintomas físicos de doenças ao invés de depressivos. Em uma tentativa de resolver este problema, Beck e colegas desenvolveram o *Beck Depression Inventory for Primary Care* (BDI-PC), uma curta escala de medida que consiste em sete itens da BDI-II considerados independentes da função física que são: tristeza, pessimismo, fracasso passado, auto-estima, autocrítica, pensamentos ou desejos suicidas e perda de interesse. Diferentemente da BDI, a BDI-PC produz apenas o resultado binário de "sem depressão"ou "com depressão"para usuários que pontuam acima de 4. Embora planejado como um instrumento de investigação e não como ferramenta diagnóstica o BDI-PC pode ser ocasionalmente utilizado para obter um rápido diagnóstico, de acordo com Gorenstein *et al.* [2011].

2.1.5 Depressão medida por escalas: estado da arte em estudos recentes

Cabe aqui ressaltar estudos relevantes quanto a métrica da depressão por meio de escalas validadas no contexto científico no mundo e no Brasil.

Conforme Sauer *et al.* [2013] a depressão é um dos transtornos de humor mais relevantes clinicamente. Por esse motivo, muitos instrumentos de avaliação foram desenvolvidos para medi-la no intuito de auxiliar o diagnóstico. Nesse sentido, o BDI é um dos instrumentos com maior frequência de uso, sendo muito aplicado em estudos no mundo e no Brasil.

Ahrari *et al.* [2013] ao realizar um estudo para determinar a gravidade dos sintomas da depressão em pacientes vítima de queimaduras e avaliar o efeito sobre esses fatores e a ocorrência da depressão no Irã, usando o inventário BDI-II, com 184 voluntários: 58 com sintomas de depressão leve, 52 moderados e 74 severa, concluiu que o BDI é um instrumento altamente confiável para triagem e validação da depressão. Foi possível com esta pesquisa destacar uma relação significativa entre os sintomas de depressão e idade, sexo e nível educacional dos voluntários, além de ter sido descoberta a necessidade do diagnóstico e tratamento com terapias psicofarmacológicas e comportamentais no hospital, devendo o tratamento ser continuado até a reabilitação.

Em outro momento Manian *et al.* [2013] analisou as dimensões fatoriais do BDI-II, naturalmente e economicamente, em uma amostra de mulheres em situações de pós-parto, avaliando a contribuição relativa. Embora a diferenciação entre os sintomas da depressão e as alterações fisiológicas relativas ao pós-parto possam ser difíceis de serem detectadas, os achados desse estudo afirmam que: mais pesquisas são necessárias para entender melhor a relação das dimensões somáticas e outras dimensões com a sintomatologia, gravidade e funcionamento da depressão.

Já Oliveira *et al.* [2014] realizou um estudo comparativo para triagem de depressão em pessoas com epilepsia com a utilização dos instrumentos Neurological Disorders Depression Inventory for Epilepsy (NDDI-E), Hospital Anxiety and Depression Scale (HADS-D) e Beck Depression Inventory II (BDIII), detectando que os três instrumentos avaliados têm utilidade clínica na triagem da depressão em pessoas com epilepsia. Sendo que, o NDDI-E e o HADS-D são instrumentos mais ágeis neste caso, e, o BDI-II se mostrou mais robusto, mais por exigir maior tempo para aplicação dificulta a sua utilização por clínicos sem tanta disponibilidade de tempo.

Com o interesse em validar uma Escala de Depressão Pós-Natal de Edimburgo e analisar sua estrutura fatorial, Loscalzo *et al.* [2015] realizou seu estudo utilizando duas amostras: a primeira composta de 334 pais onde 39 eram deprimidos; e, a segunda composta por 102 pais, dos quais 22 eram deprimidos. A partir da aplicação da versão italiana do EPDS, do IBDI-II e do Centro de Estudos Epidemiológicos Depression Scale não foi possível detectar a depressão, mas foi identificado um estado de angústia, incluindo sintomas depressivos, ansiedade e infelicidade.

Para Schutt *et al.* [2016] a triagem de sintomas depressivos é importante quando se avaliam candidatos à cirurgia bariátrica. O BDI-II e o Patient Health Questionnaire-9 (PHQ-9) são dois instrumentos de triagem de depressão amplamente utilizados. E, ao avaliar a semelhança no desempenho desses dois instrumentos descobriu que os escores do PHQ-9 e do BDI-II nos pacientes que pretendem realizar a cirurgia bariátrica estão estreitamente correlacionados. Concluindo que estes resultados apoiam o uso do BDI-II como uma alternativa viável para o PHQ-9 para triagem de pacientes que procuram a cirurgia bariátrica.

Diante da observação dos estudos descritos, consegue-se inferir que o BDI-II é utilizado nas mais diversas áreas de pesquisa, para os mais distintos objetivos específicos quando se trata da investigação de sintomas depressivos considerando o nível de depressão que cada paciente apresenta, sendo portanto o instrumento selecionado e utilizado para auxiliar no processo investigativo dos sintomas depressivos em usuários na rede social Facebook, como parte investigativa dessa pesquisa.

2.2 Tecnologia da informação e comunicação a serviço da saúde: a contribuição das redes sociais para levatamento de informações

As Tecnologias da Informação e Comunicação (TICs) são tecnologias que têm o computador e a internet como instrumentos principais, e seu uso para a prestação de cuidados na área da saúde nos países desenvolvidos tem sido amplamente explorado, com a maioria dos países desenvolvidos fazendo grandes progressos, segundo Dery *et al.* [2016]. As TICs oferecem oportunidades para indivíduos, profissionais da área médica e profissionais de saúde obterem informações, se comunicarem com outros profissionais de saúde e pacientes, oferecerem suporte básico a saúde e promoverem programas preventivos de saúde, conforme Prachi [2010] .

Ainda conforme Norman e Tesser [2015], o uso de novas TICs em saúde tem crescido nas últimas décadas, com a adoção do uso de emails e mídias sociais, contribuindo para a área da produção do conhecimento em redes, ampliando assim os canais de comunicação para acesso aos serviços de saúde.

As mídias sociais vêm sendo utilizadas em diversos domínios como forma de impulsionar o fluxo de dados e informações, a exemplo das redes sociais que possuem alguns quesitos em comum como o compartilhamento de informações, conhecimentos e interesses podendo servir como bases de dados para pesquisa científica, por exemplo.

2.2.1 As redes sociais como banco de dados secundários para a pesquisa científica

Cabe conceituar que rede social é uma estrutura social composta por pessoas ou organizações conectadas por um ou vários tipos de relação, que compartilham valores e objetivos comuns. E estas costumam reunir uma motivação comum e podem se manifestar de diferentes formas. Portanto o termo rede social pode significar criar relacionamentos com as pessoas compartilhando objetivos em comum, onde elas não necessariamente, devem estar conectadas à internet para fazer parte de uma rede social.

As principais são as redes comunitárias estabelecidas em bairros ou cidades, em geral tendo a finalidade de reunir os interesses comuns dos habitantes, melhorar a situação do local ou prover outros benefícios, já as redes profissionais, prática conhecida como *networking*, procura fortalecer a rede de contatos de um indivíduo, visando futuros ganhos pessoais ou profissionais, segundo Duarte *et al.* [2007].

Trazendo o conceito para o universo *online*, os Sites de Redes Sociais (SNSs), (ou redes sociais *online*) tais como Facebook, VK, Google+, MySpace, Twitter, Linkedin entre outros, são serviços *online*, mais especificamente, plataformas ou sites que concentram-se em construir e refletir redes sociais e interações entre as pessoas, que por exemplo, compartilham interesses e/ou atividades, como conversar, jogar, entre outras funções, conforme expõe Pittman e Reich [2016].

Já as mídias sociais, muitas vezes utilizadas equivocadamente como SNSs, diferem na sua conceitualização, para Kaplan e Haenlein [2010] mídias sociais são: "*Um grupo de aplicações para Internet construídas com base nos fundamentos ideológicos e tecnológicos da Web 2.0*[3]*, e que permitem a criação e troca de Conteúdo Gerado pelo Utilizador (UCG)*". Ou seja, mídia social é o ambiente *online* onde as informações podem ser compartilhadas, como é o exemplo de um *blog*. Nesse aspecto, um SNSs é uma parte da mídia social, em que o Facebook, por exemplo, é tanto uma rede quanto uma mídia social.

Os SNSs, conforme Ellison e Boyd [2013], particularmente são definidos como serviços *web* onde as pessoas podem: (1) construir um perfil público ou parcialmente público dentro de um sistema limitado, (2) definir uma lista de usuários com quem estabelecem uma conexão, e (3) ver a sua lista de conexões e aquelas feitas por outros usuários dentro do sistema.

Atualmente não existe nenhuma classificação das diferentes redes sociais que tenha sido aprovado por unanimidade. No entanto, de acordo com as diferentes características dos SNSs, incluindo os tipos de conteúdo gerado pelo usuário e os tipos de relacionamentos que são permitidos entre os usuários, é possível agrupá-los e organizar sua diversidade, de acordo com Fersini *et al.* [2017].

De acordo com os tipos de conteúdo gerado pelo usuário, os SNSs podem ser divididos em várias categorias, e que segundo Ferrandina e Zarriello [2014], podem ser representados por 3 grandes grupos: os microblogs como o Twitter, as redes sociais com base em conteúdo como o Youtube e as as redes sociais dirigidas à perfis, como o Facebook.

Essa última categoria com enfoque nos perfis dos usuários, o Facebook foi o primeiro recurso sobre o círculo social das pessoas (amigos, família, etc.), onde compartilham conteúdo sobre suas vidas privadas, interesses pessoais, e atividades, exprimindo o desejo de se expressar e se comunicar com seus contatos, conforme Fersini *et al.* [2017].

Dependendo das especificidades dos SNSs, existem diferentes possibilidades para que os usuários possam se expressar e se comunicar uns com os outros. Em cada plataforma de rede social, as pessoas têm várias possibilidades de interagir, e há diferentes tipos de dados que podem ser coletados (por exemplo, textos, vídeos, fotos). Há algumas diferenças fundamentais entre estas fontes, e uma compreensão exata do que elas são pode ajudar a definir formas mais eficientes para analisar a informação que elas contêm, em conformidade com Fersini *et al.* [2017].

Dentre os SNSs que podem ser utilizados como fonte de dados a fim de obter informações relevantes está o Facebook, sendo atualmente um dos mais populares do mundo, e com maior quantidade de usuários.

[3] Forma em que a World Wide Web é utilizada, sendo que o conteúdo e as aplicações são continuamente modificadas por todos os usuários de forma participativa e colaborativa. Kaplan e Haenlein [2010]

2.2.2 O Facebook e suas contribuições no campo acadêmico

O Site de Rede Social (SNS) Facebook foi criado em 2004 por Mark Zuckerberg, e atualmente é um das mais populares do mundo com 1,59 bilhões de usuários. O Brasil é o terceiro país em número de usuários, com 99 milhões de contas ativas, significando que a cada dez brasileiros que têm acesso à internet, oito tem uma conta no Facebook, onde a cada 60 segundos são postados 510.000 comentários e 293.000 atualizações de *status*. Facebook [2017]. Portanto concentra uma grande quantidade de dados que podem ser coletados para os mais diferentes objetivos.

Conforme Russell [2013], o Facebook é o coração da *web* social, visto que mais da metade dos seus 1,5 bilhões de usuários estão ativos, e a cada segundo realizando tarefas como: atualização de *status*, postagem de fotos, troca de mensagens, conversa em tempo real, realização de *check-ins* em locais físicos, jogos, compras, entre outras.

Enquanto o Twitter, por exemplo, apresenta um modelo de relacionamento de amizade assimétrico aberto e dependente de outros usuários, sem qualquer consentimento particular, o modelo do Facebook é simétrico e requer um acordo mútuo entre os usuários para que seja possível visualizar as interações e atividades uns dos outros.

O Facebook armazena os dados de cada usuário em diversas classes de dados referidas como **itens**, por exemplo, no tocante aos itens públicos do usuário, *public_profile*, possui os atributos: *id, cover, name, first_name, last_name, age_range, link, gender, locale, picture, timezone* e *updated_time*, já no que se refere às atividades do usuário, alguns itens que os contém podem ser os *user_actions_books, user_photos, user_likes* e *user_posts*.

Do ponto de vista da mineração de redes sociais, a quantidade de dados que o Facebook armazena sobre os usuários, grupos e produtos é bastante encorajadora na descoberta do conhecimento. E sua *Application Programming Interface* (API) facilita a criação de ferramentas de auxílio no acesso a essas informações. Por outro lado, esse acesso é acompanhado por uma grande responsabilidade e o Facebook preparou um conjunto de controles de privacidade *online* por meio de sua Política de Dados[4] para ajudar a proteger seus usuários da especulação.

Portanto a única maneira para obter acesso aos dados não públicos na conta de um usuário do Facebook é registrando um aplicativo (*app*) e usando-o como condição de entrada para a sua plataforma de desenvolvimento. Além disso, os únicos dados que estarão disponíveis para um *app* são os itens que o usuário tenha expressamente permitido o acesso concedendo essa permissão ao efetuar o *login* no *app*.

Além da permissão cedida pelo usuário, o *app* deverá solicitar algumas autorizações para seus itens, e cada permissão tem seu próprio conjunto de requisitos. Todas as permissões, exceto a padrão *public_profile*, exigem a habilitação do login no OAuth[5] do cliente para o aplicativo. Algumas permissões não exigem análise, porém a maioria delas requer uma análise. Qualquer utilização dessas permissões está sujeita às Políticas da Plataforma do Facebook[6] e a sua própria política de privacidade. Alguns exemplos de itens que necessitam de análise são: *ads_management, pages_messaging, publish_actions, read_insights, user_events, user_likes* e *user_posts*. (Fonte Facebook, 2017).

O documento da Política da Plataforma do Facebook, fornece os direitos e responsabilidades para todos os usuários do Facebook, bem como o conjunto de regras a serem seguidas pelos desenvolvedores de aplicativos no Facebook.

A coleta dos dados em uma SNS pode ser realizada por diversos métodos. Para pesquisas acadêmicas, segundo Rieder [2013], três formas de coleta são mais utilizadas: (1) acesso direto aos servidores, (2) acesso por meio de APIs e (3) acesso via crawlers. Para acessar a plataforma Facebook o acesso via API pode ser empregada. Na plataforma Facebook também é possível ter acesso a diversos produtos para desenvolvedores, incluindo o desenvolvimento de jogos, login em páginas à partir do Facebook, plugins sociais, e principalmente o desenvolvimento de *apps*.

Como mencionado anteriormente, para a realização da coleta de dados no Facebook é necessário

[4] Política de privacidade dos dados do Facebook, https://www.facebook.com/about/privacy/

[5] O OAuth é um protocolo de autorização para API's web voltado a permitir que aplicações clientes acessem um recurso protegido em nome de um usuário.

[6] Política da Plataforma do Facebook, https://developers.facebook.com/policy/

o uso de um *app* desenvolvido sob os critérios estabelecidos pela plataforma, incluindo configurações como nome, domínio, contato, termos de serviço, tokens de acesso, política de privacidade, envio para aprovação de itens, entre outras. Facebook [2017].

O que garante grande interação de informação entre o aplicativo e os perfis dos usuários é a *Graph API*, um serviço do tipo *Representational State Transfer* (RESTful) que retorna arquivos do tipo *JavaScript Object Notation* (JSON). O processo é feito através do envio de uma solicitação *Hypertext Transfer Protocol* (HTTP) para iniciar a conexão com a rede social, o que permite executar métodos GET para coletar dados que poderão ser posteriormente analisados. Facebook [2017].

A grande quantidade de dados que o Facebook disponibiliza em sua plataforma e a possibilidade de obtê-los por intermédio de suas APIs fornecem boas oportunidades por meio da criação de aplicativos úteis nas mais diversas áreas e para os mais variados propósitos, inclusive na área acadêmica para fins de pesquisa científica, segundo Mining the social web

Como já mencionado, o Facebook armazena os dados em formato JSON, onde há um atributo (rótulo) e um valor atrelado a ele (denominado pares atributo/valor). O trecho de um arquivo do Facebook em formato JSON exposto a seguir mostra essa relação.

```
2  "posts": [  // abre o array posts
        {   // inicia um objeto
           "message": "Sem mais...",    // atributo message
           "id": "1104080929631843_1492837554089510",
           "_id": ObjectId("590bc2f6239ac8004300d285")
        } ,  // termina um objeto

    {
           "message": "Sa sa saricando...",
           "story": "Maricy shared Anapaula's photo.",
           "created_time": ISODate("2017-03-01T01:58:44.000+0000"),
           "id": "1104080929631843_1430397880333478",
           "_id": ObjectId("590bc2f6239ac8004300d29f")
        }
    ]
```

No trecho de arquivo JSON, pode-se observar alguns dos dados do item user_posts, representado pelo array de objetos chamado posts, onde as aberturas e fechamentos de chaves delimitam cada um dos objetos podendo conter os atributos message, story, created_time e ids.

2.2.3 Tecnologia, redes sociais e depressão: estado da arte para diagnóstico e tratamento da depressão com a utilização das tecnologias

O progresso na terapia para a depressão não é apenas relativo ao avanço tecnológico da terapia por meio do computador e on-line, mas acontece de diversas formas e para os mais variados fins, podendo auxiliar pela informação sobre a doença a fim de minimizar o preconceito e aumentar a aceitação e ajuda no tratamento, segundo Vilela e Sato [2012].

A tecnologia oferece soluções nas esferas dos aspectos biológicos, psicológicos, educacionais e ambientais da depressão. Nesse contexto, destacam-se três estudos que demonstram esse avanço.

Cuijpers *et al.* [2015],] relata à partir dos resultados de seu estudo que as intervenções respaldadas na internet são eficazes no tratamento da depressão, ressaltando que a mesma está crescendo e há muitas possibilidades para todos os tipos de intervenções e aplicações inovadoras. Conclui afirmando que não deve haver dúvida de que essa área terá um impacto grande e duradouro no campo da assistência em saúde mental, no processo terapêutico e nas relações paciente-terapeuta e, portanto, mudará consideravelmente esse campo. Estudos robustos baseados em evidências devem acompanhar, no entanto, novos tratamentos de depressão com respaldo na tecnologia antes de serem implementados nos cuidados de rotina.

Já o estudo de Callan *et al.* [2017], na mesma linha de raciocínio preocupado com o tratamento de pessoas com depressão comprova que as tecnologias de computadores e dispositivos móveis podem oferecer soluções para o fornecimento de terapias para essas pessoas. Os autores ao destacarem as tecnologias disponíveis como: terapia cognitivo-comportamental assistida por computador, autoajuda por intermédio da web, grupos de suporte de autoajuda na internet, intervenções psicoterapêuticas móveis, exercícios aprimorados por

tecnologia de biosensores, alcançaram aceitação significativa pelos pacientes e atuam como coadjuvantes no tratamento da depressão.

Por outro lado, Zhao *et al.* [2017], consegue à partir de seu estudo evidenciar no período de 2004 a 2014 uma gama diversificada de TICs usadas para apoiar os programas psicoeducacionais, sendo que a maioria das intervenções utilizaram os websites como principal meio de disseminação e relatou maior uso de ferramentas de comunicação em comparação com outras abordagens eficazes, como jogos de adaptação ou tecnologias interativas, vídeos e ferramentas de auto monitoramento. Afirmando que quanto maior o uso dessas ferramentas de comunicação maior a adesão ao tratamento, apontando que futuros estudos experimentais podem ajudar a descobrir os efeitos dos recursos de tecnologia e revelar novas maneiras de auxiliar na intervenção médica.

Como denotado, existe uma infinidade de recursos tecnológicos sendo utilizados em prol da saúde, e, a cada dia mais pesquisas e soluções sendo testadas na busca de melhorias na área da Ciência da Saúde, em destaque na área da saúde mental. Quando se trata da depressão, o cenário não é diferente, especialmente no que tange a detectar mais precocemente novas situações de depressão ou de risco de suicídio, tratar as situações consoantes a sua gravidade e apoiar os tratamentos com o auxílio da tecnologia da informação, geralmente por meio da internet, tem trazido resultados bastante positivos, segundo Jamil [2017]

Neste sentido, destaca-se que os Sites de Redes Sociais (SNSs) se tornaram grandes fontes de dados que podem ser transformados em informações relevantes nesta seara. Existem vários trabalhos relevantes na área de análise de comportamento do usuário na rede social Facebook, auxiliados por modelos de identificação de patologias, a exemplo do estudo de Ortigosa *et al.* [2013], onde afere a personalidade do usuário no Facebook auxiliado pelo modelo Alternative Big Five, classificando a personalidade de acordo com cinco traços.

Oeldorf-Hirsch e Sundar [2015] comprova que o Facebook ao possibilitar aos usuários a veiculação de notícias constrói uma teia de relações sociais onde o autor da postagem atua como formador de opinião. Havendo uma diferenciação entre o ato de postar e a significação que os amigos dão para a postagem à partir de seus comentários sobre o assunto ou por apenas compartilhar ou curtir a informação. Conclui afirmando que essa descoberta demonstra que é possível os indivíduos atuarem como fontes de informações, provocando no outro um comportamento engajado sobre eventos atuais, ou seja, o Facebook promove uma articulação social capaz de mobilizar a sociedade frente a determinado conteúdo postado.

Ortigosa *et al.* [2014] comprovou por meio de seu estudo que é possível prever a personalidade do usuário por meio do acesso e das informações constantes no Facebook, o que pode contribuir para as tecnologias assistivas, e-learning, e-commerce, sistemas de saúde ou recomendação. O aplicativo por ele criado possibilita prever a personalidade a partir de parâmetros relacionados às interações do usuário, como o número de amigos ou o número de postes de parede, sendo que esta classificador apresentou alto nível de precisão.

Estudos como os de Kojouri [2015], Ozan e Pazek [2016] e Bojmela *et al.* [2016] ocuparam-se em medir quanto a interação do usuário com o Facebook alcançando resultados que hoje explicam o novo paradigma tecnológico, onde as comunicações interpessoais se estabelecem em maior número pelas redes sociais. Sendo possível por meio de experimentos, como demonstrados pelos autores citados caracterizar os usuários a partir de múltiplas variáveis dicotômicas e categóricas, mensurar o grau de satisfação de vida e os motivos que levam os usuários a utilizar-se das redes sociais para sua comunicação bem como os efeitos desta opção no campo social.

Nesta mesma linha, buscando medir a intensidade do uso do Facebook, Orosz *et al.* [2016] criou a Escala de Intensidade Multidimensional do Facebook (MFIS) validando-o a partir dos resultados encontrados pois o questionário foi capaz de diferenciar de forma confiável os diferentes aspectos da intensidade de uso do Facebook.

Já Kim e Yang [2017] comprovou por meio da análise de conteúdo das mensagens postadas e curtidas no Facebook que os comportamentos que envolvem esta rotina na rede social é carregado de comportamentos afetivos e ou cognitivos. Onde as mensagens geram diferentes comportamentos. Recursos sensoriais e visuais levaram a comentários racionais e interativos, sensitivos, visuais e racionais para compartilhar. O que demonstra o efeito das redes sociais sobre o comportamento das pessoas, ou a possibilidade de verbalizar o interno de maneira interativa.

Nos estudos de Celebi [2015] ficaram comprovados que a motivação dos jovens em relação à publicidade na Internet e publicidade no Facebook foi a utilidade interpessoal, sendo que os resultados para ambos estudos apontaram que a qualidade de vida, a influência dos colegas e o tempo de estrutura predizem significativamente o uso dos recursos de comunicação, com destaque ao sentimento de segurança e privacidade no uso do Facebook, o que contribui para o aumento da manifestação de pensamentos nos comentários deixados a cada postagem.

Em Niu *et al.* [2018], com base na teoria da comparação social, o estudo teve como objetivo investigar a associação entre o uso de SRS (Qzone) dos adolescentes chineses frente a depressão, bem como o papel mediador da comparação social negativa e o papel moderador da autoestima, chegando-se a conclusão de

que a comparação social negativa pode ser um fator-chave e um mecanismo que explica a associação positiva entre o uso do SNS e a depressão, enquanto a autoestima pode proteger os adolescentes do desfecho adverso do uso do SNS.

Interessado em demonstrar os motivos do uso e a atualização do Facebook Marshall *et al.* [2015] conseguiu uma associação entre o comportamento e os conteúdos das postagens: usuários extrovertidos atualizam sobre suas atividades sociais e vida cotidiana; usuários com baixa autoestima atualizam seus status sobre parceiros românticos; usuários com alto grau de conscientização discorrem sobre seus filhos; e, usuários narcisistas usam o Facebook para chamar a atenção e atualizam sobre suas realizações, sua dieta e rotina de exercícios. Percebe-se que foi possível traçar um reconhecimento comportamental dos usuários, o que reflete na afirmativa das redes sociais como fonte de levantamento de dados ecléticos.

A segunda série de estudos aqui apresentados estão diretamente relacionados com a possibilidade de se extrair dados do Facebook para caracterizar a personalidade dos usuários, a partir de diferentes variáveis e base de dados para estudo.

Shen *et al.* [2015] e Garcia e Sikström [2014] partindo respectivamente da análise dos estilos de escrita e número de curtidas; e, análise semântica de frequência das atualizações do status, comprovaram a existência da correlação entre a atuação do usuário no Facebook e traços de personalidade relacionados ao neuroticismo, à extroversão, a psicopatia, ao narcisismo e ao maquiavelismo. Enfatizando um comportamento socialmente malévolo, como autopromoção, frieza emocional, duplicidade e agressividade presentes nas participações e atualizações.

Já Eşkisu *et al.* [2017] examinou a relação entre os fatores de diferenças individuais e variáveis relacionadas ao Facebook, como a frequência e a finalidade com que os estudantes universitários usam o Facebook, sugerindo que pessoas com baixa autoestima e alto narcisismo podem usar o Facebook para compensar ou regular sua autoimagem e indicar a importância das diferenças individuais. Fazendo uso do modelo Big Five para a estruturação da personalidade.

Kuo e Tang [2014] e Lee *et al.* [2014] de maneira semelhante estudaram traços de personalidade humana a partir do uso do Facebook pelos usuários relativo a sua frequência. Os resultados do primeiro estudo apontaram que usuários extrovertidos e acessíveis gostam de socializar-se no Facebook por intermédio de fotos e sobre a vida real. Já as com baixa estabilidade emocional apenas para socializar-se. O segundo estudo comprovou que a frequência de comparação social das pessoas no Facebook, destacando uma associação positiva entre a frequência de comparação social e a frequência de um sentimento negativo de comparação.

Com ênfase nos fenômenos psicológicos relacionados ao Facebook Bodroža e Jovanović [2016] após aplicar seu questionário abrangente para captar processos psicológicos mais profundos que ocorrem nesta rede social, associou que o uso compensatório e viciante do Facebook relaciona-se a traços de personalidade que indicam baixa adaptabilidade social, enquanto a auto apresentação no Facebook contribui ainda mais para o processo da não adaptação. Comprovando que a utilização da escala Aspectos Psico-Sociais do Facebook Uso (PSAFU) é viável para exames detalhados das experiências dos usuários.

Blachioa *et al.* [2016] têm-se centrado na investigação da ligação potencial entre os SNSs e problemas de saúde mental. Particularmente entre o uso do Facebook, autoestima e satisfação com a vida, usando a Bergen Facebook Addiction Scale (BFAS), Escala de Intensidade do Facebook (FBI), Escala de Autoestima de Rosenberg (SES) e Escala de Satisfação com a Vida (SWLS). Os resultados do estudo mostram que os usuários do Facebook diferem estatisticamente na autoestima e satisfação com a vida. A dependência do Facebook está relacionada com menor autoestima. O vício na utilização da rede também foi negativamente relacionado com a satisfação com a vida.

Enquanto Jung *et al.* [2017] se ocupou em estudar a relação entre o uso do Facebook e o bem-estar psicológico (PWB) em adultos jovens, Carvalho e Pianowski [2017] buscou investigar em que nível há evidências da relação entre informações observáveis do perfil do Facebook e traços de personalidade patológica. Ambos estudos possibilitaram a demonstração de níveis de satisfação positivos e negativos na vida dos usuários, bem como traços de narcisismo na relação com o número de amigos no Facebook.

Lin e Utz [2015] ao explorar os resultados emocionais ao ler um post no Facebook e examinar o papel dos laços afetivos na previsão de felicidade e inveja, concluiu que as emoções positivas são mais prevalentes do que emoções negativas enquanto o usuário navega no Facebook. Além disso, o laço afetivo está positivamente associado ao sentimento de felicidade e à inveja benigna, enquanto a inveja maligna é independente do laço após a leitura de um post com conteúdo positivo no Facebook.

Smith *et al.* [2017] buscou explorar os tópicos de linguagem correlacionados com a frequência no uso de mídia social em um grupo de usuários e avaliar as diferenças na quantidade de postagens em indivíduos com diferentes diagnósticos de doenças, determindo se os pacientes poderiam prever com precisão seus próprios níveis de engajamento. E, concluiu que que postam anúncios de alta frequência foram os mais propensos a postar sobre a saúde e a ter um diagnóstico de depressão. Embora existam diversas pesquisas que contribuem significativamente para a área, a proposta do desenvolvimento do modelo proposto está sendo motivada pela

insuficiência de pesquisas tanto envolvendo dados oriundos de likes e posts na rede social Facebook, quanto da utilização de um modelo de escala de identificação de patologias específicas e adaptada à realidade brasileira, como é o caso do DBI-II, para identificação de depressão.

Por fim, o estudo de Abella *et al.* [2017] analisou a influência das redes sociais na relação entre solidão e depressão na população idosa na Espanha, onde conclui que o tipo e o tamanho das redes sociais têm um papel importante na relação entre solidão e depressão. Considerando que o aumento da interação social pode ser mais benéfico do que estratégias baseadas na melhoria da cognição social mal adaptativa na solidão, para reduzir a prevalência de depressão entre adultos idosos espanhóis.

Em resumo, diversos são os atributos que podem suportar as análises de comportamento do usuário no Facebook, a exemplo das curtidas e postagens que conforme Ryan e Xenos [2011], e Kim [2016], são as características mais utilizadas e também onde os usuários passam a maior parte do tempo, respectivamente. Respaldados pelos trabalhos selecionados e descritos, pode-se afirmar que os mais variados tipos de comportamento estão sendo extraídos de informações provenientes dos SNSs, incluindo a rede social Facebook.

Tais comportamentos estão sendo apoiados por algum tipo de escala de investigação específica a cada patologia, como é o caso do FBI, DBI, SES, entre outros, atestando a necessidade de uma escala específica para a investigação de comportamentos depressivos, a exemplo da escala BDI-II, utilizada nessa pesquisa.

2.3 Mineração de dados com ênfase na Regressão Logística: Uma ferramenta para o conhecimento

Segundo Kumar *et al.* [2009], a mineração de dados é o processo de descoberta de informações úteis em grandes bases de dados com o intuito de descobrir padrões úteis que, ao contrário, poderiam permanecer ocultos. Elas também provém a capacidade de previsão do resultado de uma observação futura.

Para Kotu e Deshpande [2015], a mineração de dados também é conhecida como descoberta de conhecimento, aprendizado de máquina e análise preditiva, no entanto, cada termo tem uma conotação ligeiramente diferente, dependendo do contexto. Witten *et al.* [2017], acrescenta que mineração de dados é um tópico prático e envolve aprender de forma prática, não de forma teórica, e que o processo pode ser automático ou semi-automático, sendo que os padrões descobertos devem ser significativos.

Witten *et al.* [2017], acrescenta que mineração de dados é um tópico prático e envolve aprender de forma prática, não de forma teórica, e que o processo pode ser automático ou semi-automático, sendo que os padrões descobertos devem ser significativos. Para o autor o processo de mineração de dados segue o modelo de referência CRISPDM, que é o acrônimo de Cross Industry Standard Process for Data Mining, e envolve conhecer o que se quer alcançar, seguindo as etapas relativas ao entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e desenvolvimento.

Kotu e Deshpande [2015], apresentam, além do processo CRISP-DM, os processos SEMMA, SAS, DMAIC, Sigma e KDD. Pelo fato de todas essas estruturas exibirem características comuns, além do fato de que um processo de mineração de dados deveria seguir um determinado conjunto de tarefas para alcançar seus objetivos, na presente pesquisa seguiremos o processo exposto em Kumar *et al.* [2009].

Kumar *et al.* [2009], em concordância com Kotu e Deshpande [2015], afirmam que a mineração de dados é uma parte integral da descoberta de conhecimento em banco de dados (*KDD Knowledge Discovery in Databases*), que é o processo geral de conversão de dados brutos em informações relevantes, conforme mostra a figura 2.2. Este processo consiste de uma série de passos de transformação, do pré-processamento dos dados até o pós-processamento dos resultados da mineração de dados.

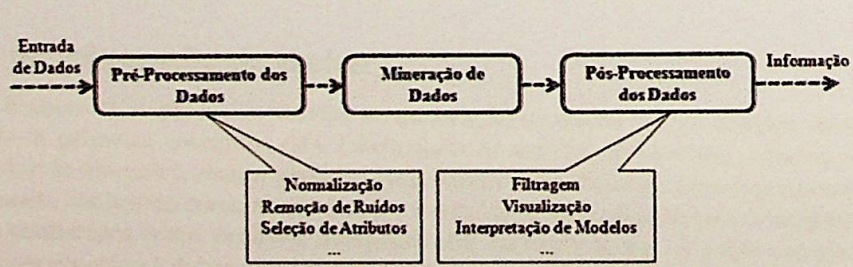


Figura 2.1: *O processo de descoberta de conhecimento em banco de dados - KDD*
Fonte: Introdução ao Data Mining, Kumar *et al.* [2009]

Ainda para Kumar *et al.* [2009], os dados de entrada podem ser armazenados em diversos formatos (arquivos simples, planilhas ou tabelas relacionais) e podem ficar em um repositório central de dados ou

serem distribuídos em múltiplos locais. O propósito do pré-processamento é transformar os dados de entrada que estão na forma bruta, em um formato apropriado para análises subsequentes. As etapas envolvidas no pré-processamento de dados, incluem a fusão de dados de múltiplas fontes, a limpeza dos dados, observações duplicadas, seleção de registros e características que sejam relevantes à tarefa de mineração de dados. Por causa das muitas formas através das quais os dados podem ser coletados e armazenados, o pré-processamento de dados talvez seja o passo mais trabalhoso e demorado no processo geral de descoberta do conhecimento. Já a etapa de pós-processamento de dados deve assegurar que apenas resultados válidos e úteis sejam incorporados nas análises, nessa fase demonstram-se os padrões de filtragem, a visualização e a interpretação das análises.

2.3.1 Mineração de dados e seu relacionamento com outras áreas de conhecimento

Segundo Han *et al.* [2012], como um domínio altamente orientado a aplicativos, a mineração de dados incorporou muitas técnicas de outros domínios, como estatísticas, aprendizado de máquina, reconhecimento de padrões, banco de dados e sistemas de armazenamento de dados, recuperação de informação, visualização, algoritmos, computação de alto desempenho, entre outros. A natureza interdisciplinar da pesquisa e o desenvolvimento da mineração de dados contribui significativamente para a sucesso da mineração de dados e suas extensas aplicações.

Kumar *et al.* [2009] afirma que o trabalho que culminou na área de mineração de dados, foi construído sobre a metodologia e algoritmos que os pesquisadores havia usado anteriormente, sendo a mineração de dados é constituída por áreas, como a (1) amostragem, estimativa e teste de hipóteses à partir de estatísticas e (2) algoritmos de busca, técnicas de modelagem e teoria de aprendizagem da inteligência artificial, reconhecimento de padrões e aprendizagem de máquina. Uma quantidade de outras áreas também desempenham papéis chave, em especial, os sistemas de banco de dados são necessários para fornecer eficiente suporte ao armazenamento, indexação e processamento de consultas. A figura 2.2 mostra o relacionamento da mineração de dados com outras áreas.

Figura 2.2: *Mineração de dados com uma união em outras disciplinas tese-fluxo*
Fonte: Introdução ao Data Mining, Kumar *et al.* [2009]

Observa-se na figura 2.2, que as áreas da estatística e da inteligência artificial ajudam a compor a área de mineração de dados.

2.3.2 Tarefas de mineração de dados

De acordo com Kumar *et al.* [2009], as tarefas de mineração de dados são geralmente divididas em duas categorias principais, a primeira categoria está relacionada às tarefas de previsão, onde o objetivo dessas tarefas é prever o valor de um determinado atributo respaldado nos valores de outros atributos. O atributo a ser previsto é geralmente conhecido como variável dependente ou alvo, enquanto que os atributos usados para fazer a previsão são conhecidos como variáveis independentes ou explicativas. A outra categoria refere-se às tarefas descritivas e seu objetivo é derivar padrões (correlações, tendências, grupos, trajetórias e anomalias), que possam resumir os relacionamentos subjacentes nos dados. As tarefas descritivas da mineração de dados são muitas vezes exploratórias em sua natureza e frequentemente requerem técnicas de pós-processamento para validar e explicar os resultados.

Em conformidade com Kotu e Deshpande [2015], de uma perspectiva histórica, existem duas classes principais das técnicas de análise preditiva: aquelas que evoluíram a partir de estatísticas como a regressão,

e aquelas que emergiram de uma mistura de estatísticas, ciências e matemática. O interesse maior nessa pesquisa está voltado às tarefa referentes a categoria preditiva, onde, segundo Kumar *et al.* [2009], a regressão é uma das suas técnicas, sendo a sua variável alvo ou dependente, aquela que está sendo avaliada, de natureza contínua.

2.3.3 Mineração de dados: Regressão Logística em foco

Em tarefas de previsão, a meta é encontrar uma função (hipótese ou modelo) à partir dos dados de treinamento que possa ser utilizada para prever um valor que caracterize um novo exemplo, com base nos valores de seus atributos de entrada, conforme Fávero e Belfiore [2017].

Um algoritmo preditivo é uma função que, dado um conjunto de dados (*dataset*) rotulados, constrói um estimador, esse rótulo toma valores em um domínio conhecido podendo ser um problema de classificação, onde o seu estimador será um "classificador", ou um problema de regressão com um estimador denominado "regressor", segundo Faceli *et al.* [2011].

A regressão é uma técnica de modelagem preditiva que possui a tarefa de aprender uma função f que mapeie cada conjunto de atributos X em uma saída de valores Y. O objetivo da regressão é encontrar uma função que possa ajustar os atributos independentes com um erro mínimo, em conformidade com Pang-Ning *et al.* [2006].

As variáveis **dependentes** que geralmente são representadas pelo eixo Y, também são conhecidas como alvo, saída, meta, resposta, rótulo, etiqueta, desfecho, fenômeno de interesse, etc, no domínio dessa pesquisa ela será referenciada como variável "dependente". Já as variáveis **independentes** representadas pelo eixo X, também chamadas de entrada, preditoras ou explicativas, nesse contexto serão referenciadas como "variáveis independentes ou explicativas".

O conjunto de técnicas de regressão é muito provavelmente o mais utilizado em análises de dados que procuram entender a relação entre o comportamento de determinado fenômeno e o comportamento de uma ou mais variáveis independentes. Até aproximadamente o início do século XX, os modelos lineares que envolviam a distribuição normal praticamente dominaram o cenário da modelagem de dados. Entretanto, a partir do período entre guerras, começam a surgir modelos para fazer frente a situações em que as modelagens lineares normais não se adequavam satisfatoriamente. Todos esses modelos acabaram sendo consolidados, do ponto de vista teórico e conceitual, por meio do trabalho de Nelder e Wedderburn [1972], em que foram definidos os *Generalized Linear Models* (GLM) ou Modelos Lineares Generalizados.

Para Casella *et al.* [2002], a definição de um GLM é descrita por uma relação entre a média de uma variável independente e uma variável dependente. Segundo a abordagem mostrada em Everitt [2012], variáveis dependentes Y são estabelecidas assim que as observações a serem feitas são definidas, podendo ser contínuas ou discretas, com o ajuste de diferentes distribuições, com médias μ_i, isto é, $E(Y) = \mu_i, i = 1, ..., n$.

O GLM, segundo Fávero e Belfiore [2017], é definido da seguinte forma:

$$\eta_i = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + ... + \beta_k.X_{ki} \tag{2.1}$$

Em que η é conhecido por função de ligação canônica, α representa a constante, $\beta_j (j = 1, 2, ..., k)$ são os coeficientes de cada variável explicativa e correspondem aos parâmetros a serem estimados, X_j são as variáveis explicativas (métricas ou *dummies*) e os subscritos i representam cada uma das observações da amostra em análise ($i = 1, 2, ..., k$), em que n é o tamanho da amostra. As demais equações seguirão o padrão exposto em Fávero e Belfiore [2017].

A tabela 2.1 relaciona cada caso particular dos GLM, com a característica da variável dependente, a sua distribuição e a respectiva função de ligação canônica.

Modelo de Regressão	Característica da Variável Dependente	Distribuição	Função de Ligação Canônica (η)
Linear	Quantitativa	Normal	$\hat{Y}$
Com transformação Box-Cox	Quantitativa	Normal após transformação	$\frac{Y^{\lambda}-1}{\lambda}$
Logística Binária	Qualitativa com 2 Categorias*(Dummy)*	Bernoulli	$ln\left(\frac{p}{1-p}\right)$
Logística Multinomial	Qualitativa M (M>2) categorias	Binomial	$ln\left(\frac{p_m}{1-p_m}\right)$
Poisson	Quantitativa com valores inteiros e não negativos (Dados de Contagem)	Poisson	$ln(\lambda)$
Binomial Negativo	Quantitativa com valores inteiros e não negativos (Dados de Contagem)	Poisson-Gama	$ln(u)$

Tabela 2.1: ***Tabela modelo para servir de exemplo***
Fonte: Análise de dados Fávero e Belfiore [2017]

A tabela 2.1, relaciona cada caso particular dos modelos lineares generalizados com a característica da variável dependente, a sua distribuição e a respectiva função de ligação canônica para uma variável dependente Y, que representa o fenômeno em estudo. A seguir 3 dos modelos apresentados serão especificados:

Modelo de regressão linear

$$\hat{Y}_i = a_i + \beta_1.X_{1i} + \beta_2.X_{2i} + ... + \beta_k.X_{ki} \tag{2.2}$$

Em que $\hat{Y}$, é o valor esperado da variável Y.

Modelo de regressão logística binária

$$\left(\frac{p_i}{1-p_i}\right) = a + \beta_1.X_{1i} + \beta_2.X_{2i} + ... + \beta_k.X_{ki} \tag{2.3}$$

Em que p é a probabilidade de ocorrência do evento de interesse definido por $Y = 1$, sendo a variável dependente *dummy*.

Modelo de regressão logística multinomial

$$\left(\frac{p_{im}}{1-p_{im}}\right) = a_m + \beta_{1m}.X_{1i} + \beta_{2m}.X_{2i} + ... + \beta_{km}.X_{ki} \tag{2.4}$$

Em que p_m ($m = 0,1,...,\ M - 1$) é a probabilidade de ocorrência de cada uma das M categorias da variável dependente Y.

Referente às variáveis *dummy*, exemplificadas anteriormente quando da definição do GLM, de acordo com Fávero e Belfiore [2017], a determinação do número de variáveis necessárias para a investigação de um fenômeno não é direta e igual ao número de variáveis utilizadas para medir os respectivos atributos, entretanto o procedimento para determinar o número de variáveis independentes cujos dados estejam em escalas qualitativas é diferente.

Kantardzic [2006] aponta que uma variável categórica com dois valores pode ser convertida em uma variável numérica binária com dois valores, 0 ou 1, e uma variável categórica com n valores pode ser convertida em n variáveis numéricas binárias, ou seja, uma variável binária para cada valor categórico. Essas variáveis categóricas codificadas são conhecidas como variáveis *dummy*. Por exemplo, se a variável cor dos olhos tiver quatro valores: preto, azul, verde e castanho, eles podem ser codificados com quatro dígitos binários, por

exemplo: preto 1000, azul 0100, verde 0010 e castanho 0001.

Ainda conforme Fávero e Belfiore [2017], não é possível simplesmente atribuir valores a cada uma das categorias da variável qualitativa, porque isso seria uma falha, denominada por ponderação arbitrária, pela suposição de que as diferenças na variável dependente seriam previamente conhecidas e de magnitudes iguais as diferenças dos valores atribuídos a cada uma das categorias da variável independente qualitativa, para eliminar esse problema, deve-se recorrer ao artifício das variáveis *dummy* ou binárias, que devem ser utilizadas quando se deseja estudar a relação entre o comportamento de determinada variável independente qualitativa e o fenômeno em questão, representado pela variável dependente.

Esse capítulo abordará os métodos de regressão à partir da regressão logística, contemplando a binária e multinomial e com alguns conceitos adjacentes à regressão logística longitudinal para dados em painel.

Antes da abordagem referente a regressão logística, alguns pressupostos serão apresentados, sendo eles a **multicolinearidade**, a **heterocedasticidade** e a **autocorrelação**.

Referente a **multicolinearidade**, conforme Fávero e Belfiore [2017] e Wayne e Cross [2013], o problema da multicolinearidade ocorre quando há correlações muito elevadas entre as variáveis independentes. Fávero e Belfiore [2017] afirma ainda que uma das principais causas da multicolinearidade é a existência de variáveis que apresentam a mesma tendência durante alguns períodos.

De acordo com Vasconcellos e Alves [2000], a existência da multicolinearidade tem impacto direto no cálculo da matriz $X'X$, podendo haver uma correlação perfeita, muito alta, porém não perfeita e uma correlação baixa.

Para Wayne e Cross [2013], como os dados estão correlacionados, não é possível encontrar soluções exclusivas para um determinado modelo. A solução menos complexa para a multicolinearidade é calcular as correlações entre todas as variáveis independentes e reter apenas as variáveis que não são altamente correlacionadas. Uma regra prática conservadora para remover a redundância no conjunto de dados é eliminar variáveis relacionadas a outras com um coeficiente de correlação significativo acima de 0,7.

Ainda conforme Fávero e Belfiore [2017], o primeiro e mais simples método para o diagnóstico da multicolinearidade refere-se à identificação de altas correlações entre as variáveis independentes por meio da análise da matriz de correlação simples. Se por um lado este método apresenta uma grande facilidade de aplicação, por outro não consegue identificar eventuais relações existentes entre mais de duas variáveis simultaneamente. O segundo método, menos utilizado, diz respeito ao estudo do determinante da matriz $X'X$, pois os valores $\det X'X$ muito baixos podem indicar a presença de altas correlações entre as variáveis independentes, prejudicando a análise estatística t e o último método a ser citado é o diagnóstico de multicolinearidade elaborado por meio da estimação de regressões auxiliares.

Segundo Vasconcellos e Alves [2000], a partir da expressão $Y_i = \alpha + b_1.X_{1i} + b_2.X_{2i} + b_k.X_{ki} + u_i$, podem ser estimadas regressões, do modo que:

$$\begin{aligned} X_{1i} &= \alpha + b_1.X_{2i} + b_2.X_{3i} + b_{k-1}.X_{ki} + u_i \\ X_{2i} &= \alpha + b_1.X_{1i} + b_2.X_{3i} + b_{k-1}.X_{ki} + u_i \\ &\vdots \\ X_{ki} &= \alpha + b_1.X_{1i} + b_2.X_{2i} + b_{k-1}.X_{k-1i} + u_i \end{aligned} \tag{2.5}$$

e, para cada uma delas haverá um R_k^2, e se um ou mais desses R_k^2 auxiliares for elevado, considera-se a existência de multicolinearidade. Definindo assim, as estatísticas *Tolerance* e *Variance Inflation Factor* (VIF), como:

$$Tolerance = 1 - R_k^2 \tag{2.6}$$

$$VIF = \frac{1}{Tolerance} \tag{2.7}$$

Assim sendo, se a *Tolerance* for muito baixa e, consequentemente a estatística VIF for alta, haverá um indício de problemas de multicolinearidade.

Fávero e Belfiore [2017] ainda declara, que enquanto muitos autores afirmam que problemas de multicolinearidade surgem com valores de VIF acima de 10, pode-se perceber que um valor de VIF igual a 4 resulta em uma *Tolerance* de 0,25 em um R_k^2 de 0,75 para determinada regressão auxiliar, o que representa um percentual relativamente elevado de variância compartilhada entre determinada variável explicativa e as demais.

Finalizando com a afirmação de David *et al.* [2001], de que a instabilidade nos parâmetros estimados é

um problema se esses valores forem o foco de interesse, ou seja, se o interesse recair em descobrir qual das variáveis é mais importante no modelo. No entanto, normalmente não será um problema, uma vez que a precisão preditiva está relacionada a diferença nos vetores que podem ser produzidos por pequenas variações dos dados, todos esses vetores levarão à predições similares para a maioria dos vetores x_k.

No que tange a **heterocedasticidade**, conforme Fávero e Belfiore [2017], a distribuição de probabilidades de cada termo aleatório em cada $Y = A + B$ é tal que todas as distribuições devem apresentar a mesma variância ou seja devem ser homocedásticas. Segundo Greene [2012], os erros de especificação quanto à forma funcional ou quanto a omissão de variável relevante podem gerar termos de erro heterocedásticos no modelo.

Para David *et al.* [2001], embora a regressão múltipla seja uma técnica muito poderosa e amplamente utilizada, algumas das suposições podem ser consideradas restritivas. A suposição de que a variância da distribuição é a mesma em cada vetor X, é muitas vezes inadequada. Esse pressuposto de variâncias iguais é chamado de homocedasticidade, enquanto o inverso é a heterocedasticidade.

Relacionado a **autocorrelação**, segundo Gujarati e Porter [2011b] o termo autocorrelação pode ser definido como a correlação entre resíduos de séries temporais em diferentes pontos no tempo, porém, um problema comum na aplicação de métodso de de regressão em séries temporais é a violação da premissa de que os resíduos, isto é, as diferenças entre valores previstos na regressão e valores observados, são independentes.

Alguns métodos podem ajudar na identificação da autocorrelação como o teste de Durbin-Watson, outro exemplo é o teste de Cochrane-Orcutt associado à transformação de Prais-Winsten, sendo o objetivo dos processos auto-regressivos solucionar o problema da violação de premissa de independência dos resíduos, apresentado em Gujarati e Porter [2010]. A equação básica usada para representar um esquema autoregressivo de primeira ordem é $Y_t = \beta 0 + \beta_1 X_t + u_t$ onde $u_t = \rho u_{t-1} + \varepsilon_t$. Se ρ for zero, significa que não há autocorrelação dos resíduos. Caso contrário, ρ representa a correlação dos termos de erro ao longo do tempo, conforme mostra Gujarati [2000].

Para finalizar, a explanação relativa aos pressupostos apresentados, pode-se ressaltar que existem os problemas, causas e consequências tanto na **multicolinearidade** como na **heterocedasticidade** e **autocorrelação**, porém existem testes diagnósticos e possíveis soluções para cada um desses problemas por meio de métodos estatísticos adequados.

Prosseguindo na abordagem referente a regressão logística, conforme Wooldridge [2013], em modelos de regressão linear (simples ou múltipla), a variável dependente Y, é uma variável aleatória de natureza contínua. Porém , em alguns casos, essa variável é qualitativa e, portanto, representada por duas ou mais categorias, isto é, admite dois ou mais valores. Neste caso, o método dos mínimos quadrados não oferece estimadores admissíveis. No entanto, por meio da regressão logística, pode-se obter uma boa aproximação, já que esta permite o uso de um modelo de regressão para calcular ou prever a probabilidade de um evento específico. Intensificando a sentença, Fávero e Belfiore [2017] afirma que, diferentemente da tradicional técnica de regressão, estimada pelo método dos métodos de mínimos quadrados ordinários (MQO), em que a variável dependente apresenta-se de forma contínua, as técnicas de regressão logística são utilizadas quando o fenômeno a ser estudado apresenta-se de forma qualitativa e, portanto, representado por uma ou mais variáveis *dummy*, dependendo da quantidade de possibilidades de respostas (categorias) dessa variável dependente.

Conforme Hoffmann [2016], a regressão logística é o membro mais importante (e provavelmente mais usado) da classe de modelos GLM. Ao contrário da regressão linear, a regressão logística pode prever diretamente valores restritos ao intervalo (0,1), como probabilidades, sendo também uma boa primeira escolha para problemas de classificação binária.

Os modelos de regressão logística podem ser subdivididos em duas categorias, binária e multinomial.

2.3.4 Regressão Logística Binária

O modelo de regressão logística binária tem como objetivo principal estudar a probabilidade de ocorrência de um evento definido por Y, que se apresenta na forma qualitativa dicotômica $Y = 1$, para descrever a ocorrência do evento de interesse e $Y = 0$, para descrever a ocorrência de um não evento, esse tipo de variável dependente é chamado de variável Bernoulli, que será explicado à frente, com base no comportamento das variáveis independentes. Dessa forma pode-se definir um vetor de variáveis explicativas, com os respectivos parâmetros estimados da seguinte forma, segundo Fávero e Belfiore [2017].

$$Z_i = \alpha + \beta_1 . X_{1i} + \beta_2 . X_{2i} + ... + \beta_k . X_{ki} \tag{2.8}$$

Em que Z é o *logit*, e α representa a constante, $\beta_j (j = 1, 2, ..., k)$ são os parâmetros estimados de cada variável explicativa, X_j são as variáveis explicativas (métricas ou *dummies*) e o subscrito i representa cada observação da amostra ($i = 1, 2, ..., n$ em que n é o tamanho da amostra). O objetivo nesse momento é definir a expressão da probabilidade p_i de ocorrência do evento de interesse para cada observação, em função do

logit Z, que são os parâmetros estimados para cada variável explicativa. Para tanto, o conceito de chance de ocorrência, também denominado *odds*, é definido da seguinte forma:

$$chance(odds)_{yi=1} = \frac{p_i}{1-p_i} \tag{2.9}$$

A regressão logística binária define o *logit* Z como o logaritmo natural da chance de modo que:

$$ln(chance_{y_i=1}) = Z_i \tag{2.10}$$

de onde vem que:

$$ln\left(\frac{p_i}{1-p_i}\right) = Z_i \tag{2.11}$$

No entanto, o *logit* não tem conceito igual ao de probabilidade, o que pode gerar conclusões probabilísticas equivocadas. Tendo como objetivo definir uma expressão para a probabilidade de ocorrência do evento em estudo em função do logit, p_i pode matematicamente ser isolado à partir da expressão 2.11 da seguinte maneira:

$$\frac{p_i}{1-p_i} = e^{Z_i} \tag{2.12}$$

$$p_i = (1-p_i).e^{Z_i} \tag{2.13}$$

$$p_i.(1+e^{Z_i}) = e^{Z_i} \tag{2.14}$$

E, portanto, tem-se :

$$p_i = 1 - \frac{e^{Z_i}}{1+e^{Z_i}} = \frac{1}{e^{Z_i}} \tag{2.15}$$

Que é a probabilidade de ocorrência do evento.

$$1 - p_i = 1 - \frac{e^{Z_i}}{1+e^{Z_i}} = \frac{1}{e^{Z_i}} \tag{2.16}$$

$$P(y=1|X) = \frac{1}{1+e^{-}Z_i} \tag{2.17}$$

E a probabilidade de ocorrência do não evento.

$$P(y=0|X) = 1 - \frac{1}{1+e^{-}Z_i} = \frac{exp^{-}(Z)}{1+e^{-}Z_i} \tag{2.18}$$

Em particular, a regressão logística assume que o $logit(y)$ é linear nos valores de x. Como a regressão linear, a regressão logística encontrará os melhores coeficientes para prever y, incluindo encontrar combinações e cancelamentos vantajosos quando as entradas são correlacionadas, o que a regressão logística binária estima não são os valores previstos da variável dependente mas sim a probabilidade de ocorrência do evento em estudo para cada observação. Os parâmetros das duas funções de ligação são estimados de forma iterativa pelo método da máxima verossimilhança, pois são transformações das distribuições acumuladas, segundo Hoffmann [2016].

Em relação ao **Método de Máxima Verossimilhança**, de acordo com Fávero e Belfiore [2017], este consiste em minimizar a soma de quadrados das diferenças entre os valores observados e os previstos. Na regressão não linear o método da máxima verossimilhança é utilizado de forma iterativa para que sejam encontradas as estimativas mais prováveis dos parâmetros. Ao invés de minimizar os desvios dos quadrados, a regressão não linear maximiza a probabilidade de que um evento ocorra.

Segundo Gujarati e Porter [2011a], um método de estimação preciso e com algumas propriedades teóricas mais fortes que as do método dos mínimos quadrados ordinários é o da máxima verossimilhança (MV), se for considerada a distribuição de u_i normal, os estimadores de máxima verossimilhança e de mínimos quadrados ordinários dos coeficientes de regressão, os β, serão idênticos e isso será válido tanto para as regressões simples quanto para as múltiplas. Para elucidar melhor os conceitos pertinentes a estimação por máxima verossimilhança, pode-se fazer uso de um exemplo, adaptado de Fávero e Belfiore [2017], onde a variável dependente será qualitativa e dicotômica. No exemplo, um professor explorou os efeitos de determinadas variáveis explicativas sobre o tempo de deslocamento de um grupo de acadêmicos até a universidade, onde

o interesse estava em investigar se essas variáveis explicativas influenciam a probabilidade de um acadêmico chegar atrasado à aula, ou seja, o fenômeno apresenta somente duas categorias (chegar ou não chegar atrasado) e o evento de interesse refere-se a chegar atrasado.

Sendo assim, o professor elaborou uma pesquisa com 100 acadêmicos, questionando se cada um deles chegou ou não atrasado naquele dia, questionando também sobre a distância percorrida no trajeto em km, representada pela variável *dist*, o número de semáforos pelo qual cada um passou, representado pela variável *sem*, o período em que foi realizado o trajeto (manhã ou tarde), representado pela variável *per* e como cada um se considera em termos de perfil ao volante (calmo, moderado ou agressivo). Para a variável dependente, como o evento de interesse referia-se a chegar atrasado, esta categoria apresentava valores iguais a 1, ficando a categoria não chegar atrasado com valores iguais a 0, a categoria de referência da variável correspondente ao período do dia foi *tarde*, ou seja, as células da base de dados com essa categoria assumiram valores iguais a 0, ficando as células com a categoria *manhã* com valores iguais a 1. A variável perfil ao volante foi transformada em duas *dummies*, onde $perfil2$ representou a categoria *moderado* e $perfil3$ representou a categoria *agressivo*, e a categoria *calmo* foi definida como sendo a referência. O *logit*, cujos parâmetros serão estimados, é definido pela expressão:

$$Z_i = \alpha + \beta_1.dist_i + \beta_2.sem_i + \beta_3.per_i + \beta_4.perfil2_i + \beta_5.perfil3_i \tag{2.19}$$

e a probabilidade estimada de que um determinado acadêmico chegue atrasado pode ser escrito da seguinte forma:

$$p_i = \frac{1}{1 + e^{-(\alpha+\beta_1.dist_i+\beta_2.sem_i+\beta_3.per_i+\beta_4.perfil2_i+\beta_5.perfil3_i)}} \tag{2.20}$$

Como não faz sentido definir o tempo de erro para cada observação, dado que a variável dependente apresenta-se na forma dicotômica, não há como estimar os parâmetros da equação de probabilidade por meio da minimização da somatória dos quadrados dos resíduos, como nas técnicas tradicionais de regressão, como já exposto anteriormente. Nesse caso, a função de verossimilhança foi utilizada, à partir da qual foi elaborada a estimação por máxima verossimilhança. Quanto a **Distribuição de Bernoulli**, segundo Hoffmann [2016], os ensaios de Bernoulli têm apenas dois resultados, que por conveniência são rotulados como 0 para um resultado "mal-sucedido" e 1 para um resultado "bem-sucedido".

Na regressão logística binária, segundo Fávero e Belfiore [2017] a variável dependente segue uma distribuição de Bernoulli, ou seja, o fato de determinada ação ou não evento de interesse, pode ser considerado como um ensaio de Bernoulli, em que a probabilidade de ocorrência do evento é p e a probabilidade de ocorrência de um não evento é $1 - p$. Descreve-se a probabilidade de ocorrência de Y_i podendo Y_i ser igual a 1 ou a zero e a fórmula é dada por:

$$P(Y_i) = (P_i)^{Y_i} . (1 - P_i)^{1-Y_i} \tag{2.21}$$

A função de verossimilhança, para uma amostra com n observações pode ser definida como:

$$L = \prod_{i=1}^{n} \left[(P_i)^{Y_i} . (1 - P_i)^{1-Y_i} \right] \tag{2.22}$$

de onde vem, com base nas expressões 2.15 e 2.18, que:

$$L = \prod_{i=1}^{n} \left[\left(\frac{e^{Z_i}}{1 + e^{Z_i}} \right)^{Y_i} \cdot \left(\frac{1}{1 + e^{Z_i}} \right)^{1-Y_i} \right] \tag{2.23}$$

Como é mais conveniente trabalhar com o logaritmo da função de verossimilhança, pode-se chegar a função *log likelihood function*:

$$LL \sum_{i=1}^{n} \left[(Y_i).ln \left(\frac{e^{Z_i}}{1 + e^{Z_i}} \right) \right] + \left[(1 - Y_i).ln \left(\frac{e^{Z_i}}{1 + e^{Z_i}} \right) \right] \tag{2.24}$$

A fim de estimar os parâmetros $\alpha, \beta_1, \beta_2, ..., \beta_k$, apresenta-se a função-objetivo (*maximum likelihood estimation*):

$$LL \sum_{i=1}^{n} \left[(Y_i).ln \left(\frac{e^{Z_i}}{1 + e^{Z_i}} \right) \right] + \left[(1 - Y_i).ln \left(\frac{e^{Z_i}}{1 + e^{Z_i}} \right) \right] = max \tag{2.25}$$

continuando com o exemplo do professor, após o cálculo do somatório do logaritmo da função de ve-

rossimilhança, chega-se ao valor de -69,31472, e o valor máximo da sua somatória é $LL_{max} = -29,06568$, gerando as seguintes estimativas dos parâmetros:

$\alpha = -30,202$
$\beta = 0,220$
$\beta = 2,767$
$\beta = -3,653$
$\beta = 1,346$
$\beta = 2,914$

$$p_i = \frac{1}{1 + e^{-(a+\beta_1.dist_i+\beta_2.sem_i+\beta_3.per_i+\beta_4.perfil2_i+\beta_5.perfil3_i)}} \tag{2.26}$$

Após a estimação por verossimilhança, pode-se partir para o estudo da significância estatística geral do modelo obtido, e das significâncias estatísticas dos parâmetros.

Em se tratando da significância geral do modelo e dos parâmetros da regressão logística binária, destaca-se que existem outros métodos para gerar o R^2, onde todos medem quão bem o modelo é ajustado e seus coeficientes variam entre 0 e 1. o R^2 de McFadden, esse coeficiente é igual a 1 menos a razão entre a probabilidade do modelo ajustado e a probabilidade do modelo independente; o R^2 de Cox e Snell, este coeficiente é igual a 1 menos a razão entre a probabilidade do modelo ajustado e a probabilidade do modelo independente elevado à potência 2/Sw, onde Sw é a soma dos pesos; o R^2 de Nagelkerke, este coeficiente é igual à razão do R^2 de Cox e Snell, dividido por 1 menos a probabilidade do modelo independente elevado à potência 2/Sw.

Em Greene [2012], encontra-se a seguinte afirmação: "uma medida de adequação originalmente proposta para modelos de seleção discreta no ano de 1974 por McFadden, que surpreendentemente ganhou grande aceitação em toda a literatura empírica, foi o índice de razão de verossimilhança, que passou a ser conhecido como o *pseudo* R^2."

Gujarati e Porter [2011a] afirma que a medida convencional de qualidade de ajuste, R^2, não é particularmente significativa em modelos de regressão binária, porém medidas semelhantes à R^2, chamadas de *pseudo*R^2, estão disponíveis, uma discussão acerca de algumas dessas medidas podem ser encontradas em detalhes em Long [1997].

Sua utilização na presente pesquisa será delimitada ao *pseudo* R^2 de *McFadden* exposto em Fávero e Belfiore [2017].

Continuando no exemplo do professor, contido em Fávero e Belfiore [2017], como a variável dependente é qualitativa, não faz sentido a discussão do percentual de sua variância, que é explicado pelas variáveis preditoras, ou seja, em modelos de regressão logística não há um coeficiente de ajuste R^2 como nos modelos tradicionais de regressão estimados pelo método de mínimos quadrados ordinários, portanto o coeficiente *pseudo* R^2 de *McFadden*, será utilizado, cuja expressão é representada por:

$$pseudoR^2 = \frac{-2.(LL_0 - LL_{max})}{-2.(LL_0} \tag{2.27}$$

E com base na expressão 2.27, obtem-se:

$$pseudo\,R^2 = \frac{-2.(-67,68585) - (-29,06568)}{-2.(-67,68585)} = 0,5706 \tag{2.28}$$

Fávero e Belfiore [2017], ainda afirma que o *pseudo* R^2 pode ser usado como um indicador de desempenho, porém o melhor indicador de desempenho de um modelo de regressão logística binária refere-se a eficiência global do modelo, definida com base na determinação *cutoff*. Teste X^2 (Hipóteses nula e alternativa) Após o cálculo da significância do modelo, pode-se calcular a significância estatística geral do modelo por meio do teste X^2, que tem suas hipóteses nula e alternativa, representadas por:

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
$H_1 : existe\,pelo\,menos\,um\,\beta_j \neq 0$

O cálculo do teste X^2 é o mais adequado para métodos estimados pelo método de máxima verossimilhança, como os modelos de regressão logística e sua equação é representada por:

$$X^2 = -2.(LL_0 - LL_{max}) \tag{2.29}$$

Voltando ao exemplo, tem-se:

$$X^2_{5g.l.} = -2.[-67,68585 - (-29,06568)] = 77,2403 \tag{2.30}$$

Tanto para os graus de liberdade (variáveis independentes) quanto para o nível de significância, o valor

é 5. Assim, o valor na tabela X^2 é 11,070. Dessa forma como $X^2_{calc} = 77,2403 > X^2_c = 11,070$ rejeita-se a hipótese nula, de que todos os parâmetros $\beta_j (j = 1, 2, ..., k)$, sejam iguais a zero. Isso significa que ao menos uma variável X é estatisticamente significante para explicar a probabilidade de ocorrência do evento em estudo, gerando assim um modelo de regressão logística estatisticamente significantes para fins de previsão. Análogo ao teste F, o teste de X^2 avalia a significância conjunta das variáveis independentes, não sendo possível saber quais dessas variáveis são significantes. Nesse caso, a estatística $z\,de\,Wald$ será importante para fornecer a significância estatística de cada parâmetro a ser considerado no modelo.

Segundo Stubbe e Coleman [2014], a estatística de Wald é o quadrado da divisão do coeficiente pelo respectivo erro padrão e segue uma distribuição qui-quadrado, sendo que quanto maior for o valor estatístico, mais forte será a influência da variável no modelo.

Para Fávero e Belfiore [2017], a nomenclatura z significa que a distribuição é a normal padrão. As expressões para o cálculo das estatísticas $z\,de\,Wald$ de cada parâmetro α e β_j, são representadas por:

$$z_\alpha = \frac{\alpha}{s.e(\alpha)} \tag{2.31}$$

$$z_{\beta_j} = \frac{B_j}{s.e(\beta_j)} \tag{2.32}$$

em que $s.e$ é o erro padrão (*standard error*) de cada parâmetro em análise. Continuando com o exemplo, os valores $s.e$ de cada parâmetro, são:

$s.e\ (\alpha) = -9,981$
$s.e\ \beta_1 = 0,110$
$s.e\ \beta_2 = 0,922$
$s.e\ \beta_3 = 0,878$
$s.e\ \beta_4 = 0,748$
$s.e\ \beta_5 = 1,179$

Calculadas as estimativas dos parâmetros, tem-se:

$$z_\alpha = \frac{\alpha}{s.e(\alpha)} = \frac{-30,202}{9,981} = -3,026 \tag{2.33}$$

$$z_{\beta_1} = \frac{B_1}{s.e(\beta_1)} = \frac{0,220}{0,110} = 2,000 \tag{2.34}$$

$$\vdots$$

$$z_{\beta_5} = \frac{B_5}{s.e(\beta_5)} = \frac{2,914}{1,179} = 2,472 \tag{2.35}$$

Após a obtenção das estatísticas $z\,de\,Wald$, a tabela de distribuição da curva normal padrão para obtenção dos valores críticos, a um dado nível de significância, pode ser utilizada. A rejeição ou não da hipótese nula pelos testes podem ser verificadas.

Quanto ao **Modelo de Regressão Probit**, conforme Fávero e Belfiore [2017], modelos de regressão probit, cujo nome se refere à contração de *probability unit*, podem ser utilizados alternativamente aos modelos de regressão logística binária, para os casos em que a curva de probabilidade de ocorrência de determinado evento ajusta-se mais adequadamente a função densidade de probabilidade acumulada da distribuição normal padrão.

Os modelos de regressão *probit* são muito utilizados para a compreensão de relação dose-resposta, quando a respectiva curva de probabilidade de ocorrência do evento de interesse, inicialmente representado por uma variável binária, seguir uma função sigmóide.

Para Greene [2012], o *probit* é um tipo de modelo de classificação binária, cujo objetivo do modelo é estimar a probabilidade de que uma observação com características particulares caia em uma das categorias específicas, classificando as observações com base em suas probabilidades previstas.

Continuando com Fávero e Belfiore [2017], no modelo *probit* a variável dependente segue uma distribuição de Bernoulli e, portanto, a expressão da função objetivo (logaritmo da função de verossimilhança) tem por intuito estimar os parâmetros $\alpha, \beta_1, \beta_2, ..., \beta_k$ de determinado modelo de regressão *probit* para um modelo de regressão logística binária, dada por:

$$LL \sum_{i=1}^{n} [(Y_i).ln\,(p_i)] + [(1 - Y_i).ln\,(p_i)] = max \tag{2.36}$$

O que varia, portanto, entre os modelos de regressão logística binária e os modelos de regressão *probit* é a expressão das probabilidades de ocorrência do evento de interesse p_i. Conforme visto, na regressão logística binária de p_i, que apresenta distribuição logística é dada por:

$$p_i = \frac{1}{1+e^{-Z_i}} \frac{1}{1+e^{-(a+\beta_1.X_{1i}+\beta_2.X_{2i}+...+\beta_k.X_{ki})}} \tag{2.37}$$

Para a regressão *probit*, a expressão das probabilidades de ocorrência do evento de interesse, que apresentam distribuição normal padrão acumulada, pode ser expressa por:

$$p_i = \Phi(Z_i) = \Phi(a+\beta_1.X_{1i}+\beta_2.X_{2i}+...+\beta_k.X_{ki}) \tag{2.38}$$

Em que Φ representa a própria função densidade de probabilidade acumulada da distribuição normal padrão. Nesse sentido, a expressão 2.38 pode ser escrita conforme segue:

$$p_i = \int_{-\infty}^{Z_i} \frac{1}{\sqrt{2.\pi}}.e^{\left(\frac{1}{2}.Z^2\right)}dz \tag{2.39}$$

para facilidade de cálculo pode ser reescrita da seguinte maneira:

$$p_i = \frac{1}{2} + \frac{1}{2}.\left(1-e^{\frac{-2.Z_i^2}{\pi}}\right)^{\frac{1}{2}} = para\ Z \geq 0 \tag{2.40}$$

$$p_i = 1 - \left[\frac{1}{2} + \frac{1}{2}.\left(1-e^{\frac{-2.Z_i^2}{\pi}}\right)^{\frac{1}{2}}\right] = para\ Z < 0 \tag{2.41}$$

À partir das expressões 2.37, 2.40 e 2.41, pode-se elaborar a tabela 2.2, que apresenta valores de p em função de valores de Z variando de -5 a +5, tornando possível a comparação entre as curvas logística (*logit*) e *probit* de probabilidades.

	Regressão Logit	Regressão Probit
Z_i	p_i	
-5	0,01	0,00
-4	0,02	0,00
-3	0,05	0,00
-2	0,12	0,02
-1	0,27	0,16
0	0,50	0,50
1	0,73	0,84
2	0,88	0,98
3	0,95	1,00
4	0,98	1,00
5	0,99	1,00

Tabela 2.2: *Probabilidade de ocorrência de um evento (p) em função de Z para os modelos de regressão logit e probit*

Fonte: Análise de dados Fávero e Belfiore [2017]

À partir da tabela 2.2 pode-se elaborar um gráfico de $p = f(Z)$, como apresentado na figura 2.3. Por meio deste gráfico, verifica-se que embora as probabilidades estimadas em função dos diversos valores assumidos por Z situam-se entre 0 e 1 para ambos os casos, parâmetros distintos serão estimados pelos modelos *logit* e *probit*, visto que diferentes valores de Z são necessários para que se chegue à mesma probabilidade de ocorrência do evento de interesse para determinada observação i.

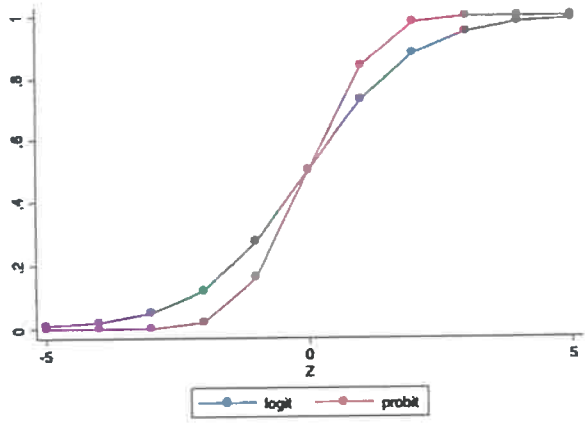


Figura 2.3: *Gráfico de $p = F(Z)$ para os modelos logit e probit*
Fonte: Elaboração própria

Conforme pode-se observar pelo gráfico da figura 2.3, as funções *logit* e *probit* não são consideravelmente distintas, principalmente para valores de Z em torno de zero, sendo que os parâmetros estimados em cada caso seguem a relação $\alpha, \beta_{logit} \approx 1,6[\alpha, \beta_{probit}]$, conforme discute Amemiya [1981].

Conforme aponta Finney [1952], a opção pela escolha do modelo *probit*, em detrimento do modelo *logit*, ocorre pela aderência da curva de probabilidades de ocorrência do evento de interesse a distribuição normal padrão acumulada e a decisão pode ser tomada com base em quatro critérios:

Modelo com mais alto valor do logaritmo da função de verossimilhança;

Modelo com maior pseudo R^2 de McFadden;

Modelo com mais alto nível de significância do teste de hosmer lemeshow (menor estatística x quadrado deste teste);

Modelo com maior área abaixo da curva ROC.

Assim, dentro desse item, foram apresentadas as funções de ligação *logit* e *probit*, as quais possuem a variável binária como dependente. O *logit* e o *probit* utilizam funções de distribuição específicas para a realização do cálculo da probabilidade, que são respectivamente, a logística e a normal.

2.3.5 Regressão Logística Multinomial

Segundo Allison [2014], a regressão logística binária é ideal quando a variável dependente tem apenas duas categorias, porém quando ela possui mais de duas categorias em alguns casos, pode ser razoável reduzir as categorias para apenas duas, mas essa estratégia envolve inevitavelmente alguma perda de informação, ou ainda ao reduzir as categorias poderia causar imprecisão do que se está tentando estudar. O modelo é chamado de *logit* multinomial, e isso ocorre porque a distribuição de probabilidade para o resultado é referente a uma variável multinomial em vez de uma distribuição binomial.

Seguindo na mesma linha, conforme Fávero e Belfiore [2017], para estimar a probabilidade de ocorrência, é necessário definr a categoria de referência. Por exemplo, uma variável dependente se apresenta na forma qualitativa com três categorias possíveis de resposta (0, 1 ou 2). Se for definida a categoria 0 como sendo a referência, tem-se duas outras possibilidades de evento em relação a essa categoria, que serão representadas pelas categorias 1 e 2, sendo assim dois vetores de variáveis explicativas serão definidos com os respectivos parâmetros estimados, gerando dois *logits*:

$$Z_{i1} = a_1 + \beta_{11}.X_{1i} + \beta_{21}.X_{2i} + ... + \beta_{k1}.X_{ki} \tag{2.42}$$

$$Z_{i2} = a_2 + \beta_{12}.X_{1i} + \beta_{22}.X_{2i} + ... + \beta_{k2}.X_{ki} \tag{2.43}$$

sendo que o número do *logit* está no subscrito de cada parâmetro a ser estimado.

A variável dependente tiver M categorias de resposta, $(M-1)$ será o número de *logits* estimados, sendo possível estimar as probabilidades de ocorrência de cada categoria a partir deles. E expressão do *logit* $Z_{i_m}(m = 0, 1, , ..., M-1)$ quando a variável assume M categorias, é:

$$Z_{i_m} = \alpha_m + \beta_{1_m}.X_{1i} + \beta_{2_m}.X_{2i} + ... + \beta_{k_m}.X_{ki} \tag{2.44}$$

onde, $Z_{i_0} = 0$ e então $e^{Z_{i_0}} = 1$.

Na seção anterior, duas categorias serviram de base para os exemplos, com apenas um *logit* Z_i. Para 3 categorias, estima-se a probabilidade de ocorrência da categoria 0 (referência) e dos dois eventos representados pelas categorias 1 e 2, tendo as seguintes expresões de probabilidade:

Probabilidade de ocorrência da categoria de referencia 0:

$$p_{i_0} = \frac{1}{1 + e^{-Z_{i_1}} + e^{-Z_{i_2}}} \tag{2.45}$$

Probabilidade de ocorrência da categoria 1:

$$p_{i_1} = \frac{e^{-Z_{i_1}}}{1 + e^{-Z_{i_1}} + e^{-Z_{i_2}}} \tag{2.46}$$

Probabilidade de ocorrência da categoria 2:

$$p_{i_2} = \frac{e^{-Z_{i_2}}}{1 + e^{-Z_{i_1}} + e^{-Z_{i_2}}} \tag{2.47}$$

Para um modelo onde a variável dependente assume M categorias de resposta, a expressão das probabilidades $p_{i_m}(m = 0, 1, , ..., M - 1)$

$$p_{i_m} = \frac{e^{Z_{im}}}{\sum_{m=0}^{m-1} e^{Z_i}} \tag{2.48}$$

Assim como na regressão logística binária na seção 2.3.4, pode-se avaliar a significância estatística geral do modelo e dos parâmetros, estimar seus intervalos de confiança a um determinado nível de significância.

Para a estimação do modelo de regressão logística multinomial por máxima verossimilhança, pode-se utilizar uma variável dependente que possui 3 valores distintos, ou seja, rótulos (*labels*) de cada uma das categorias de resposta ($M = 3$), porém agora cada categoria possui um *logit* que representa a referência (0) e outros dois que representam as categorias 1 e 2. O mesmo ocorre com as probabilidades de ocorrência de cada evento correspondente a cada categoria da variável dependente a serem estimadas. A significância estatística geral do modelo, bem como a construção dos intervalos de confiança dos parâmetros da regressão logística multinomial seguem a mesma lógica da regressão logística binária.

Genericamente, na regressão logística multinomial, em que a variável dependente segue uma distribuição binomial, uma observação i pode incidir em um determinado evento de interesse, dados M eventos possíveis, e portanto a probabilidade de ocorrência $P_{im}(m = 0, 1, ..., M - 1)$ deste específico evento pode ser escrita:

$$p(Y_{i_m}) = \prod_{m=0}^{M-1} (P_{i_m})^{Y_{im}} \tag{2.49}$$

Para uma amostra de n observações, pode-se definir a função de verossimilhança (*likelihood function*), da seguinte forma:

$$L = \prod_{i=1}^{n} \prod_{m=0}^{M-1} (P_{i_m})^{Y_{im}} \tag{2.50}$$

de onde vem, a partir da expressão 2.48, que:

$$L = \prod_{i=1}^{n} \prod_{m=0}^{M-1} \left(\frac{e^{Z_{im}}}{\sum_{m=0}^{m-1} e_{im}^{Z}} \right)^{Y_{im}} \tag{2.51}$$

Analogamente ao procedimento adotado quando do estudo da regressão logística binária, o logaritmo da função de verossimilhança será utilizado, chegando-se a função *log likelihood function*.

$$L = \sum_{i=1}^{n} \sum_{m=0}^{M-1} \left[(Y_{i_m}).ln \left(\frac{e^{Z_{im}}}{\sum_{m=0}^{m-1} e_{im}^{Z}} \right) \right] \tag{2.52}$$

Para exemplificar, tem-se hipoteticamente M categorias da variável dependente, os valores dos parâmetros dos logits Z_{im} $(m = 0, 1, ..., M - 1)$, representados pela expressão 2.44, que fazem com que o valor de LL da expressão 2.52 seja maximizado é por meio do uso de ferramentas de programação linear, com a seguinte função-objetivo:

$$L = \sum_{i=1}^{n} \sum_{m=0}^{M-1} \left[(Y_{i_m}).ln\left(\frac{e^{Z_{im}}}{\sum_{m=0}^{m-1} e_{im}^{Z}} \right) \right] = max \tag{2.53}$$

Esse exemplo é a chave central para a elaboração da estimação dos parâmetros da regressão logística multinomial por máxima verossimilhança ou *maximum likelihood estimation.*

Após a elaboração da estimação por máxima verossimilhança dos parâmetros das equações de probabilidade de ocorrência de cada uma das categorias da variável dependente, é possível elaborar a classificação das observações e definir a eficiência global do modelo de regressão logística multinomial. Diferentemente da regressão logística binária, em que a classificação é elaborada com base na defnição de um *cutoff*, na regressão logística multinomial a classificaçào de cada observação é feita com base na maior probabilidade entre as calculadas (p_{i_0}, p_{i_1} ou p_{i_2}).

Da mesma forma como apresentado na seção referente a regressão logística binária, o seguinte critério pode ser definido:

Se p-valor (ou p-value ou sig.X^2_{cal} ou Prob.X^2_{cal}) $< 0{,}05$, existe pelo menos um $\beta_{jm} \neq 0$.

Além da significância geral do modelo, a verificação da significância de cada parâmetro, por meio da análise das respectivas estatísticas z de Wald (z_c), cujas hipóteses, nula e alternativa, são para os parâmetros $\alpha_m (m = 1, 2, ..., M-1)$ e $\beta_{jm} (j = 1, 2, ..., k; m = 1, 2, ..., M-1)$, respectivamente:

$$H_0 : \alpha_m = 0$$
$$H_1 : \alpha_m \neq 0$$
$$H_0 : \beta_{jm} = 0$$
$$H_1 : \beta_{jm} \neq 0$$

Percebe-se, que os cálculos utilizam sempre as estimativas médias dos parâmetros, a próxima seção apresentará o estudo dos intervalos de confiança desses parâmetros.

Em relação a construção dos intervalos de confiança dos parâmetros do modelo de regressão logística multinomial, os intervalos de confiança dos parâmetros do modelo de regressão logística multinomial são calculados por meio da expressão:

$$\begin{matrix} \alpha_m \pm 1,96.[s.e.(\alpha_m)] \\ \beta_{jm} \pm 1,96.[s.e.(\beta_{jm})] \end{matrix} \tag{2.54}$$

Em que 1,96 é o z_c para o nível de significância de 5%. Em que segundo Field [2013], o nível de confiança é o complemento do nível de significância, isto é, um intervalo de confiança de 95% reflete um nível de significância de 0,05.

Os intervalos de confiança são usualmente estimados pelo método de Wald, cujo fundamento está na aproximação assintótica da distribuição normal, conforme exposto em Cirillo *et al.* [2010].

Pode-se definir a expressão dos intervalos de confiança das chances (it odds ou relative risk ratios) de ocorrência de cada um dos eventos representados pelo subscrito $m (m = 1, 2, ..., M-1)$ em relação a ocorrência do evento representado pela categoria 0 (referência) para cada parâmetro $\beta_{jm} (j = 1, 2, ..., k; m = 1, 2, ..., M-1)$, ao nível de confiança de 95%, da seguinte forma:

$$e^{\beta_{jm} \pm 1,96.[s.e.(\beta_{jm})]} \tag{2.55}$$

Ao expor os modelos longitudinais de regressão para dados em painel, que são o foco da presente pesquisa cabe ressaltar segundo Wooldridge [2013], que estes correspondem a um conjunto de dados de corte transversal, mais conhecido por cross-section, consiste em uma amostra de indivíduos, consumidores, empresas, cidades, estados, países ou uma variedade de outras unidades, tomada em um determinado ponto no tempo.

Às vezes, os dados de todas as unidades não correspondem precisamente ao mesmo período, já um conjunto de dados de séries temporais consiste em observações sobre uma variável ou muitas variáveis ao longo do tempo. Um conjunto de dados de painel, também conhecido por dados longitudinais, consiste em uma série de tempo para cada membro do corte transversal do conjunto de dados.

Dentro dos chamados Modelos Lineares Generalizados - *Generalized Linear Models* (GLM), estão inseridos os modelos de regressão simples e múltipla, os modelos de regressão logística e os modelos de regressão para dados de contagem, que possuem uma abordagem prioritariamente em cross section, ou seja, com exemplos de base de dados que reproduzem de certa forma, uma fotografia do momento em que são coletados os

dados. Em outras palavras, para modelos em *cross-section*, os indivíduos variam, porém o tempo é fixo. Já as bases de dados temporais, reproduzem os dados em forma de um filme da evolução temporal de determinadas variáveis, mas para um único usuário, portanto, para modelos em séries temporais, os períodos de tempo variam, mas para um único usuário, conforme Fávero e Belfiore [2017].

De acordo com Hoffmann [2016], os dados em painel, também são usados para estudar comportamentos, ou outros resultados que podem mudar com o tempo, sendo que o interesse pode ser referente, por exemplo, não apenas em, por exemplo, se os adolescentes estão envolvidos em alguma delinquência, mas também o seu nível de envolvimento durante períodos de tempo específicos. Da mesma forma, o interese em alterações nos níveis de autoestima, o envolvimento na comunidade, as pontuações nos testes ao longo do tempo ou o crescimento físico durante a infância. Portanto, os dados longitudinais permitem o estudo da dinâmica de vários resultados, o que pode ser superior ao fato de apenas examinar um resultado em um momento, como acontece com dados transversais.

Portanto os dados em painel são um tipo especial de dados combinados nos quais a mesma unidade *cross-section* é pesquisada ao longo do tempo. Ela têm uma dimensão espacial e outra temporal conforme pode ser visualizado na figura 2.4.

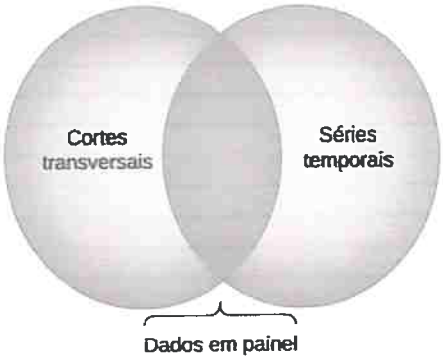


Figura 2.4: *Formação dos dados em painel*
Fonte: Elaboração própria

Na figura 2.5 observa-se, para os modelos de regressão para dados em painel, as estruturas de dados agrupados com medidas repetidas e longitudinais e a relação entre elas, o aninhamento no dados e evolução temporal com destaque para a hierarquia referente aos modelos para dados em painel.

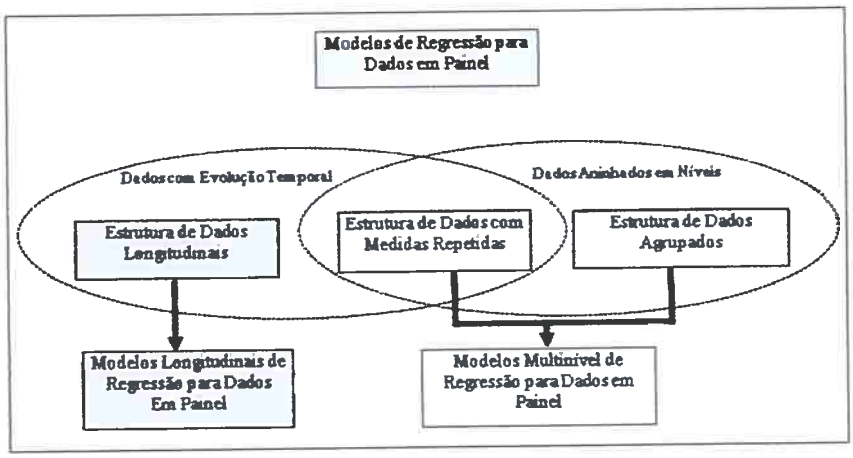


Figura 2.5: *Estrutura em modelos de regressão para dados em painel*
Fonte: Elaboração própria

Embora os dados em painel, segundo Gujarati e Porter [2011b], assumam diversas terminologias, tais como dados agrupados ou *pooled data* (*pooling* de séries temporais e observações *cross – sectional*), combinação de séries temporais e dados *cross – section* dados de micropanel, dados longitudinais (estudo temporal de uma variável ou grupo de indivíduos), análise do histórico de eventos (o movimento ao longo do tempo

dos sujeitos através de sucessivos estados ou condições), análise de *coorte*, mesmo que existam variações sutis, todos esses nomes implicam movimento ao longo do tempo em unidades $cross-sectional$, e aqui nesse contexto o nome "painel"ou *pooled* serão utilizados.

Conforme Fávero e Belfiore [2017], os modelos longitudinais de regressão tem como objetivo principal estudar o comportamento de determinada variável dependente quantitativa ou qualitativa Y, que representa o fenômeno de interesse, fundamentado no comportamento de variáveis explicativas cujas alterações podem ocorrer tanto entre indivíduos na mesma *cross-section* quanto ao longo do tempo.

A tabela 2.3 apresenta o modelo de um banco de dados longitudinal.

Observação	**Indivíduo i**	**Período t**	Y_{it}	X_{1it}	X_{2it}	...	X_{kit}
1	1	t_{11}	$Y_{1t_{11}}$	$X_{1t_{11}}$	$X_{21t_{11}}$		$X_{k1t_{11}}$
2	1	t_{21}	$Y_{1t_{21}}$	$X_{1t_{21}}$	$X_{21t_{21}}$		$X_{k1t_{21}}$
	⋮	⋮	⋮	⋮	⋮		⋮
	1	T_1	Y_{1T_1}	X_{11T_1}	X_{21T_1}		X_{k1T_1}
	2	t_{12}	$Y_{1t_{12}}$	$X_{1t_{12}}$	$X_{21t_{12}}$		$X_{k2t_{12}}$
	2	t_{22}	$Y_{1t_{22}}$	$X_{1t_{22}}$	$X_{21t_{22}}$		$X_{k2t_{22}}$
	⋮	⋮	⋮	⋮	⋮		⋮
	2	T_2	Y_{2T_2}	X_{12T_2}	X_{22T_2}	...	X_{k2T_2}
⋮	3	t_{13}	$Y_{3t_{13}}$	$X_{13t_{13}}$	$X_{23t_{13}}$		$X_{k3t_{13}}$
	3	t_{23}	$Y_{3t_{23}}$	$X_{13t_{23}}$	$X_{23t_{23}}$		$X_{k3t_{23}}$
	⋮	⋮	⋮	⋮	⋮		⋮
	3	T_3	Y_{3T_3}	X_{13T_3}	X_{23T_3}		X_{k3T_3}
	n	t_{1n}	$Y_{nt_{1n}}$	$X_{1nt_{1n}}$	$X_{2nt_{1n}}$		$X_{knt_{1n}}$
	n	t_{2n}	$Y_{nt_{2n}}$	$X_{1nt_{2n}}$	$X_{2nt_{2n}}$		$X_{knt_{2n}}$
	⋮	⋮	⋮	⋮	⋮		⋮
	n	T_n	Y_{nT_n}	X_{1nT_n}	X_{2nT_n}		X_{knT_n}

Tabela 2.3: *Modelo geral de um banco de dados longitudinal*
Fonte: Análise de dados Fávero e Belfiore [2017]

Por meio do modelo geral de banco de dados longitudinais apresentado na tabela 2.3, verifica-se que pode existir uma quantidade diferente de períodos para cada um dos n indivíduos da amostra, e que cada indivíduo apresenta dados correspondentes às variáveis Y_{it}, X_{1it}, X_{2it}, ... ,X_{kit} em cada um dos respectivos períodos de tempo. Assim, enquanto o tempo $Y_{1t_{11}}$, por exemplo, refere-se ao dado (quantitativo ou qualitativo) que assume a variável dependente Y para o indivíduo 1 no período $t = 1$, o termo $X_{22_{t_2}}$ corresponde ao valor que assume a variável explicativa X_2 para o indivíduo 2 no instante de tempo $t = T_2$ (período final para o indivíduo 2). Se $T_1 = T_2 = T_3 = T_n$, o painel será considerado **balanceado**, e a quantidade total de observações no banco de dados N será igual a $n.T$. Caso contrário, a quantidade de observações no banco de dados será igual a $\sum_i T_i$, e o painel será considerado não balanceado ou **desbalanceado**.

Ainda referente ao tema painel balanceado e desbalanceado, apresenta-se um exemplo de um estudo da teoria do investimento proposto por Y. Grunfeld, e exposto em Gujarati e Porter [2011b], onde o interesse de Grunfeld era descobrir como o investimento bruto real Y dependia do valor real da empresa X_2 e o estoque de capital real X_3, com dados de 4 empresas durante 20 anos, existindo portanto, 4 *cross section* e 20 períodos de tempo, completando 80 observações. A priori, espera-se que Y esteja positivamente relacionado com X_2 e X_3. Agrupando, ou combinando, as 80 observações, pode-se escrever a função de investimento de Grunfeld como:

$$Y_{it} = \beta_1 + \beta_2.X_{2it} + \beta_3.X_{3it} + e_{it} \tag{2.56}$$

Sendo $i = 1, 2, 3, 4$ e $t = 1, 2, ...20$, onde i é a $i - esima$ unidade $cross - sectional$ e t para o $t - esimo$ período de tempo, onde i denota o identificador $cross - section$ e t o identificador de tempo. Presume-se que haja um máximo de n unidades ou observações $cross - sectional$ e um máximo de T períodos de tempo. Se cada unidade $cross - sectional$ tem o mesmo número de observações de séries temporais, então esse painel de dados é chamado de painel balanceado. Nesse exemplo tem-se um painel balanceado, pois cada empresa da amostra tem 20 observações. Se o número de observações diferir entre os membros do painel, ele será um painel não balanceado.

Referente a quantidade de períodos contidos em diferentes bases de dados, o que caracteriza a qualidade de um painel apresentar-se balanceado ou não, diferentes modelos de dados em painel poderiam ser utilizados, dependendo do objetivo da pesquisa, à seguir será apresentado um modelo para dados em painel, que deverá nortear a presente pesquisa.

Conforme JD *et al.* [2010] o modelo geral para dados de painel é representado por:

$$Y_{it} = \beta_1 + \beta_2.X_{2it} + \beta_3.X_{3it} + \varepsilon_{it}$$

Nessa notação, o subscrito $_i$ representa os diferentes indivíduos e o subscrito $_t$ representa o período de tempo que está sendo analisado, β_0 refere-se ao parâmetro de intercepto e β_k faz alusão ao coeficiente angular correspondente a *k-ésima* variável explicativa do modelo. A forma matricial para o *i-ésimo* indivíduo pode ser representada por:

$$Y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in} \end{bmatrix} X_i = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1r} \\ 1 & x_{21} & x_{22} & \dots & x_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nr} \end{bmatrix} \beta = \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n1} & \beta_{n2} & \dots & \beta_{nm} \end{bmatrix} \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix} \tag{2.57}$$

Em que y_i e e_i são vetores de dimensão $Tx1$ e contém, respectivamente, T variáveis dependentes e os T erros. X_i é uma matriz de dimensão (K x T) com as variáveis independentes do modelo. Assim o elemento X_{kit} refere-se a *k-ésima* variável explicativa para o indivíduo i no instante de tempo t, e β_i é a matriz dos parâmetros a serem estimados.

Com o intuito de estimar os parâmetros de um modelo que considere Y_{it} em função de X_{1it}, X_{2it}, ... ,X_{kit}, a expressão geral de um modelo longitudinal de regressão, segundo Fávero e Belfiore [2017], pode ser definida da seguinte forma:

$$Y_{it} = a_i + b_1.X_{1it} + b_2.X_{2it} + ... + b_k.X_{kit} + \varepsilon_{it} \tag{2.58}$$

Em que Y representa o fenômeno em estudo, α_i representa o intercepto, $b_j (j = 1, 2, ..., k)$ são os coeficientes de cada variável, X_j são as variáveis independentes (métricas ou *dummies*) e ε representa os termos de erro idiossincrático, os subscritos i representam cada um dos indivídios da amostra em análise $(i = 1, 2, ..., n)$, em que n é a quantidade de indivíduos na amostra, e os subscritos t representam os períodos em que foram coletados os dados.

Caso a variável Y seja quantitativa, à expressão 2.58, pode ser considerada um modelo longitudinal linear de regressão, porém caso a variável Y seja qualitativa dicotômica, será considerado um modelo longitudinal logístico de regressão (modelo longitudinal não linear), considerando o último modelo, a expressão 2.58 poderá ser reescrita da seguinte maneira:

$$ln(chance_{y_{it}=1}) = a_i + \beta_1.X_{1it} + \beta_2.X_{2it} + ... + \beta_k.X_{kit} \tag{2.59}$$

Sendo a meta estimar os parâmetros α_i e $\beta_j (j = 1, 2. .... k)$, por intermédio de determinado método, para que possa ser compreendido o comportamento do fenômeno em estudo, representado pela variável dependente Y, entre indivíduos e ao longo do tempo, em função do comportamento das variáveis explicativas X_j.

Referente ao modelo geral de dados em painel, representado pela equação 2.58, faz-se necessário que sejam verificadas antes da estimação, as intensidades das variações que ocorrem temporalmente para cada indivíduo e em cada uma das *cross sections*, uma vez que, enquanto as variações temporais podem indicar a existência de mudanças no comportamento das variáveis em cada indivíduo, as variações em cada *cross section* podem indicar a existência de comportamentos discrepantes das variáveis entre indivíduos.

Dessa forma, a variação ao longo do tempo para dado indivíduo é conhecida por variação *within* e a variação entre indivíduos é chamada de variação *between*. A variação *overall* (geral), pode ser definida como sendo a discrepância que existe em determinado dado de um indivíduo no instante de tempo em relação a todos os demais dados daquela mesma variável para a base completa, que pode ser decomposta nas variações ao longo do tempo para cada indivíduo (*within*) e entre indivíduos (*between*).

De acordo com Cameron e Trivedi [2009], considerando como exemplo determinada variável X e com base em expressões de variância, pode-se definir:

Variação *within*

$$Var_{Xw} = \frac{\sum_{it} (X_{it} - \bar{X}_i)^2}{(\sum_i T_i) - 1} \tag{2.60}$$

Variação *between*

$$Var_{Xb} = \frac{\sum_{it}(X_{it} - \bar{X}_i)^2}{n-1} \tag{2.61}$$

Variação *overall*

$$Var_{Xo} = \frac{\sum_{it}(X_{it} - \bar{X}_i)^2}{(\sum_i T_i) - 1} \tag{2.62}$$

em que X_{it} representa o dado da variável X para o indivíduo i no instante de tempo t, $\bar{X}_i$ é a média da variável X para cada indivíduo i e $\bar{X}_i$ é a média geral da variável X no banco de dados. Além disso, n representa a quantidade total de indivíduos e $\sum_i T_i$ corresponde a quantidade total de observações na amostra. Se o banco de dados em painel for balanceado, pode-se substituir o termo $\sum_i T_i$ por $(n.T)$ nas expressões 2.60 e 2.62.

Segundo Hoffmann [2016], o efeito aleatório ou *between* tende a se concentrar na interceptação, sendo permitido variar aleatoriamente entre as unidades α_i, possibilitando assim, que a linha de base das variáveis de resultado possa variar entre os indivíduos. Outra característica significativa da regressão de efeitos aleatórios é que ela pode acomodar covariáveis variantes e invariantes no tempo. Esse modelos, são apropriados quando há a necessidade de examinar o que afeta as alterações em unidades individuais (por exemplo, pessoas ou organizações).

O modelo de regressão de efeitos fixos em geral pode ser representado como:

$$Y_{it} = \alpha + \beta x_{it} + \varepsilon_{it} \tag{2.63}$$

Na equação 2.63, o intercepto inclui um subscrito $_i$ que denota que existem diferentes valores de linha de base individuais para cada unidade nos dados.

Referente ao modelo de regressão de efeitos aleatórios, ainda Segundo Hoffmann [2016], ele pode ser caracterizado por:

$$Y_{it} = \alpha + \beta x_{it} + \mu_{it} + \varepsilon_{it} \tag{2.64}$$

Neste modelo, conforme mostra a equação 2.64 erro é dividido nos componentes *between* (μ_{it}) e *within* (ε_{it}). Uma suposição é que os efeitos específicos individuais não estão correlacionados com as variáveis independentes, assim, as variáveis que não variam ao longo do tempo podem ser incluídas no modelo. Como em muitos outros tipos de modelos de regressão, as variáveis omitidas não deverão afetar as variáveis incluídas no modelo, caso contrário, é omitido o viés da variável. Uma diferença fundamental entre os modelos de efeito fixo e aleatório refere-se a suposição que eles fazem sobre a correlação entre o que não é observado e as variáveis independentes no modelo.

2.3.6 Modelos Longitudinais não Lineares

Segundo Gujarati e Porter [2011a], quando o fenômeno principal sobre o qual há o interesse de estudo é representado por uma variável dependente, que apresenta dados qualitativos com valores dicotômicos, faz sentido a estimação por modelos longitudinais não lineares de regressão.

Nos **Modelos Longitudinais não Lineares** Fávero e Belfiore [2017], descreve que o fenômeno principal sobre o qual há o interesse de estudo é representado por uma variável dependente, que apresenta dados qualitativos com valores dicotômicos, faz sentido a estimação por modelos longitudinais não lineares de regressão.

A tabela 2.4, relaciona cada caso particular dos modelos longitudinais de regressão para dados em painel com a característica da variável dependente, a sua distribuição e a respectiva função de ligação canônica.

Modelo Longitudinal de Regressão para Dados em Painel	Característica da Variável Dependente	Distribuição	Função de Ligação Canônica (η)
Linear	Quantitativa	Normal	Y
Não Linear Logístico	Qualitativa com 2 Categorias *(Dummy)*	Bernoulli	$ln\left(\frac{p}{1-p}\right)$
Não Linear Poisson	Quantitativa com valores inteiros e não negativos (Dados de Contagem)	Poisson	$ln(\lambda)$
Não Linear Binomial Negativo	Quantitativa com valores inteiros e não negativos (Dados de Contagem)	Poisson-Gama	$ln(u)$

Tabela 2.4: *Modelos longitudinais de regressão para dados em painel, características da variável dependente e funções de ligação canônica*
Fonte: Análise de Dados Fávero e Belfiore [2017]

Observa-se, por meio da tabela 2.4, que para cada modelo exitem propriedades específicas, como a característica das variáveis dependentes, no presente estudo, são qualitativas, maior atenção está sendo dispensada à distribuição *Bernoulli*, a função de ligação canônica para esse modelo está inserida nas equações descritas ao longo do capítulo, para o modelo de dados em painel, a equação será apresentada no decorrer desse item.

Em se tratando da **estimação de modelos longitudinais logísticos**, a expressão geral da chance de ocorrência do evento em estudo para determinado indivíduo i em um específico instante de tempo t, representado por $X_{it=1}$, é definida de acordo como segue:

$$ln(chance_{y_{it}=1}) = a_i + \beta_1.X_{1it} + \beta_2.X_{2it} + ... + \beta_k.X_{kit} \tag{2.65}$$

que resulta na seguinte expressão de probabilidade de ocorrência do evento de interesse:

$$p_i = \frac{e^{(a_i+\beta_1.X_{1it}+\beta_2.X_{2it}+...+\beta_k.X_{kit})}}{1+e^{(a_i+\beta_1.X_{1it}+\beta_2.X_{2it}+...+\beta_k.X_{kit})}} \tag{2.66}$$

para um modelo não linear logístico

$$ln\left(\frac{p_i}{1-p_i}\right) = a_i + \beta_1.X_{1it} + \beta_2.X_{2it} + ... + \beta_k.X_{kit} \tag{2.67}$$

e que apresenta parâmetros que podem ser estimados com base na maximização do logaritmo da função de verossimilhança , cuja expressão é reproduzida a seguir para as situações em que existem dados longitudinais.

$$LL\sum_{t=1}^{T}\sum_{i=1}^{n}\left[(Y_{it}).ln\left(\frac{e^{(a_i+\beta_1.X_{1it}+\beta_2.X_{2it}+...+\beta_k.X_{kit})}}{1+e^{(a_i+\beta_1.X_{1it}+\beta_2.X_{2it}+...+\beta_k.X_{kit})}}\right)\right]+\left[(1-Y_{it}).ln\left(\frac{1}{1+e^{(a_i+\beta_1.X_{1it}+\beta_2.X_{2it}+...+\beta_k.X_{kit})}}\right)\right] \tag{2.68}$$

Analogamente aos modelos lineares, a primeira e mais simples estimação dos parâmetros de um modelo longitudinal logístico, que é elaborada por meio do método de máxima verossimilhança, é conhecida por *poolled logit*, por considerar que a base de dados seja uma grande *cross-section*. Assim como para a estimação POLS, no entanto a estimação *poolled logit* também deve considerar a existência de erros padrão robustos com agrupamento por indivíduo, a fim de que haja o controle da correção dos termos de erro para dado indivíduo ao longo do tempo conforme discutem Cameron e Trivedi [2009].

Além da tradicional estimação *poolled logit*, que gera correlações entre termos de erro iguais a zero, para dois quaisquer períodos de tempo distintos e para dado indivíduo ($p_{ts} = 0, t \neq s$), é possível que os parâmetros do modelo sejam estimados levando-se em consideração a existência de correlações diferentes de zero entre termos de erro provenientes de períodos períodos de tempo distintos. Nesse caso, a estimação mais comum é aquela em que sejam consideradas correlações iguais (diferentes de zero), entre os termos de erro para dois períodos de tempo distintos, ou seja , entre os termos de erros sejam equicorrelacionadas ($p_{ts} = p$). Esta estimação conhecida por PA (*Population Averaged Estimation*), podendo ser utilizada em modelos longitudinais logísticos e também aplicável a modelos lineares.

Ressalta-se que as estimações *poolled logit* e *PA logit* inserem-se dentro do que é conhecido por *Generalized Estimating Equations (GEE)*. Segundo Hoffmann [2016], uma vantagem dos GEEs sobre outros métodos de análise de dados longitudinais é que eles permitem diferentes estruturas de correlação serem assumidas pela

unidade within do modelo ou seja, uma estrutura de correlação pode ser imposta aos erros within da unidade e comparar diferentes tipos para chegar ao melhor ajuste do modelo.

Para os modelos lineares, os parâmetros do modelo apresentado na fórmula 2.66 podem ser estimados por efeitos fixos ou por efeitos aleatórios, levando-se em consideração que a_i seja, respectivamente, um efeito fixo ou um efeito aleatório.

Um maior aprofundamento da teoria pertinente às estimações *Pooled Logit* e PA Logit, bem como as definidas por efeitos fixos e por efeitos aleatórios, pode ser encontrado em Cameron e Trivedi [2009] e Hubbard *et al.* [2010].

Os modelos longitudinais de regressão para dados em painel são primordiais quando se deseja estudar o comportamento de determinado fenômeno, representado pela variável dependente Y, na presença de estruturas de dados agrupados, dados com medidas repetidas ou dados longitudinais. Esses modelos são cada vez mais utilizados em diversas áreas do conhecimento, visto que muitos dados de indivíduos (pessoas, empresas, municípios, estados ou países, por exemplo) estão disponíveis, não para um único instante de tempo, (uma única *cross-section*), mas em vários períodos de tempo (várias *cross-sections*, como semanas, meses, trimestres ou anos, por exemplo). Neste sentido, estimam-se modelos para o estudo de fenômenos que sofrem influência das diferenças entre os indivíduos e da própria evolução temporal e, devido ao desenvolvimento computacional dos softwares de modelagem, pode-se verificar um incremento na utilização de tais modelos, com melhores condições de investigar comportamentos e tendências em estruturas mais complexas de banco de dados, segundo Fávero e Belfiore [2017].

Além disso, segundo Baltagi [2005], os modelos longitudinais de regressão providenciam a maior quantidade de informação, maior variabilidade dos dados, menor multicolinearidade entre as variáveis, maior número de graus de liberdade e maior eficiência quanto da estimação de seus parâmetros. A inclusão da dimensão em *cross-section*, em um estudo temporal, confere maior variabilidade aos dados, na medida em que a utilização de dados agregados resulta em séries mais suaves do que as séries individuais que lhe servem de base. Esse aumento na variabilidade dos dados pode contribuir para a redução da eventual multicolinearidade existente entre as variáveis.

Para um maior aprofundamento da teoria pertinente a essas estimações recomenda-se o estudo de Cameron e Trivedi [2009] e Wooldridge [2005].

Uma tabela, que apresenta, de forma consolidada, as principais estimações dos modelos longitudinais de regressão para dados em painel, encontra-se no Apêndice I. O assunto é vasto e novos estimadores podem ser levados em consideração quando da modelagem de dados longitudinais dependendo do contexto e do objetivo de cada pesquisa.

Nesse tópico, foi explanado brevemente a teoria alusiva aos modelos de regressão, com maior ênfase aos modelos de regressão logística binária pois o modelo de regressão logística multinomial, embora seja o foco desse estudo, é uma especialidade do modelo binário. À seguir, serão apresentados alguns estudos relacionados à regressão logística, tanto multinomial na forma *cross section* quanto sob a forma longitudinal, com o objetivo de mostrar casos que utilizam modelos de regressão para predizer indícios referentes a alguma patologia por meio do comportamento apresentado e capturado em suas manifestações.

2.3.7 Mineração de dados, Regressão Logística e Depressão: estudos recentes à partir de redes sociais com ênfase na Regressão Logística Multinomial

Nesta seção alguns trabalhos envolvendo as três áreas mais relacionadas a pesquisa serão mencionados, sendo: a rede social, o local onde os usuários expressaram suas opiniões; a mineração de dados com aporte da regressão logística e a depressão como patologia selecionada.

Em um primeiro momento destaca-se a utilização da regressão logística no campo da ciência da saúde, a exemplo do estudo de Wiest *et al.* [2015], Mufudza e Erol [2016], Bertens *et al.* [2016] e, Madhu *et al.* [2014]. Sendo que nos quatro estudos concluiu-se que a regressão logística é uma ferramenta poderosa para avaliar a relação entre uma covariável ou exposição e um resultado de evento binário, e fornece a capacidade de se ajustar facilmente a possíveis fatores de confusão ao examinar associações de interesse no campo da saúde. Os demais estudos apresentados tratam diretamente da utilização da regressão logística multinomial em experimentos exitosos no campo da saúde mental.

O trabalho de Kuramoto *et al.* [2013] utilizou regressão logística multinomial, em um estudo longitudinal, para estimar a relação entre a ideação suicida e o planejamento em níveis de densidade em afro-americanos, onde as descobertas mostram que a integração social, avaliada pela densidade de redes sociais, é um importante preditor de ideação e plano de suicídio entre afro-americanos.

Em Randall *et al.* [2014] os dados sobre a ocorrência de fatores de risco demográficos, psicossociais e socioambientais foram testados por meio de regressão logística multinomial quanto à associação com ideação suicida e tentativas de suicídio, em indivíduos de Benin, revelando que o sexo feminino, a ansiedade, a

solidão, a violência física e o uso de drogas ilícitas estão associados a esses resultados.

Nierop e Germeys [2016], empregou a regressão logística multinomial para avaliar se o trauma na infância estava mais fortemente associado aos sintomas afetivos, psicóticos, ansiosos e maníacos de forma isolada do que todos os sintomas em conjunto, onde os achados revelaram que o trauma estava estreitamente associado aos sintomas de forma conjunta ao invés de isolada.

Na pesquisa de Liu [2016b], modelos de regressão logística multinomial foram usados para estimar a associação entre depressão pré-natal paterna e materna e a prematuridade alta e moderada em que os resultados mostraram que que tanto a depressão pré-natal materna inédita como a recorrente foram associadas a um risco aumentado de parto prematuro moderado.

A investigação de Kingsbury *et al.* [2018], emprega a regressão logística multinomial em um estudo longitudinal, para determinar se o número de situações adversas relacionadas ao nascimento previam trajetórias de depressão moderada-estável e/ou moderada-crescente nos vinte e sete anos subsequentes, tendo sido descoberto que essas situações adversas comuns que normalmente ocorrem não predizem trajetórias depressivas de longo prazo. Na mesma linha o estudo de Guo *et al.* [2018] o qual estimou as associações univariadas e multivariadas entre depressão maior e depressão menor com fatores de risco potenciais.

Cichowitz *et al.* [2017], avalia a saúde mental (depressão, ansiedade e transtornos por uso de álcool) no momento da iniciação da terapia antiretroviral e retenção nos cuidados durante um período de acompanhamento de seis meses. Sendo que a análise bivariada foi utilizada para procurar associações entre os dados demográficos, laboratoriais e de pesquisa e a retenção nos cuidados seis meses após a inclusão. A variável idade e demais variáveis, com valores significativos na análise bivariada, foram incluídos em uma regressão logística multivariada para a avaliação dos preditores do não comprometimento com os cuidados após a terapia, tendo como resultado que tanto a depressão como o distúrbio por álcool, estão independentemente associados ao não envolvimento com os cuidados pós terapia.

O estudo de Kimbrel *et al.* [2016] determinou o impacto de comorbidades nos sintomas depressivos e de stress em receptores de transplante cardíaco. Para avaliar a depressão foi utilizado o questionário BDI, para o stress (Perceived Stress Scale-10), para as análises estatísticas foram utilizadas a estatística descritiva, correlações de Pearson, t-testes e modelos lineares generalizados. Os resultados referentes em uma ocorrência depressão e alto nível de stress, utilizando a regressão logística multivariada foram previstos por alguns fatores independentes onde se chegou à conclusão de que a prevalência de depressão e stress severo é comum entre os receptores de transplante cardíaco.

Dentre as descobertas de Liu [2016a], pode-se destacar o estudo que desenvolveu um modelo de predição de risco de Depressão Pós AVC com base em características clínicas e sócio-psicológicas do paciente para a detecção precoce de pacientes com depressão de alto risco. A regressão logística multivariada foi utilizada para extrair os fatores de risco para depressão em um mês após o AVC ter ocorrido, onde o modelo logístico foi convertido em um modelo de árvore usando o método da árvore de decisão, este estudo forneceu um modelo de risco efetivo para Depressão Pós AVC e indicou que os fatores sócio psicológicos foram importantes fatores de risco para essa depressão.

Wutchiett e Lovasi [2017] afirma em seu estudo que a depressão clínica está associada à doença médica comórbida e à redução da adesão ao tratamento. Isto foi possível com a utilização da regressão logística para avaliar a associação da depressão ao longo da vida com a falta de assistência médica necessária no último ano, com estratificação por status de seguro de saúde e ajuste para características socioeconômicas. Onde os resultados revelaram que a depressão prévia foi correspondia a uma maior probabilidade de não ter recebido o relatório de atendimento médico necessário no ano anterior, independentemente do status do seguro de saúde, emprego, renda e dados demográficos.

Já os achados de Watabe *et al.* [2015] utilizam jogos de confiança em rede social, para avaliar as relações interpessoais do mundo real como um novo candidato a avaliações psiquiátricas, com oitenta e um estudantes universitários japoneses que responderam a um conjunto de questionários. A análise de regressão revelou que o apoio da família para os participantes do sexo masculino, e agitação subjetiva e/ou retardo para os participantes do sexo feminino, foram associados aos comportamentos de confiança dos participantes, ou seja, o apoio familiar foi negativamente associado ao comportamento cooperativo em relação aos não apoiados e as mulheres com maior agitação subjetiva confiaram menos dinheiro para os homens e mulheres atraentes, e ao contrário para mulheres atraentes com menor estatura nos relacionamentos interpessoais.

As descobertas de Ophir *et al.* [2017] utilizam-se do SNS Facebook para comparar o quadro clínico tradicional "off-line"da depressão com suas manifestações on-line e exploram características únicas da depressão on-line que são menos dominantes "off-line", avaliam até que ponto uma atualização de status contém referências à depressão, onde por meio de abordagens baseadas em teoria bottom-up, um esquema de codificação foi desenvolvido, que diferenciam entre "atualizações de status depressivos"e "atualizações de status não-depressivos". Além disso, uma análise de regressão múltipla revelou quatro características de atualização de status que predizem escores de depressão em atualização de status.

Neste contexto, fica comprovado que os modelos de regressão logística são utilizados de diversas formas e em diferentes domínios, para a previsão de alguma patologia, onde pode-se observar que os modelos apresentaram resultados positivos mostrando ser escolhas promissoras na tarefa de predição.

Capítulo 3

Marco Metodológico

A pesquisa científica para Lefchld e Barros [1991] corresponde ao exercício da inquisição, do procedimento sistemático e intensivo, com o objetivo de descobrir e interpretar fatos reais. Neste sentido, este capítulo apresenta quanto ao tipo de pesquisa, os métodos e técnicas, a população e a amostra que desenham este estudo, bem como procedeu-se a coleta e a análise dos dados.

3.1 Tipos de Pesquisa

Classifica-se esta pesquisa quanto à sua natureza, sua abordagem, seus objetivos e procedimentos.

Em relação à natureza classifica-se como pesquisa aplicada por "gerar conhecimentos para aplicação prática e dirigida à solução de problemas e interesses sociais" Silveira e Córdova [2009]. Sobretudo por buscar alternativas para subsidiar a área da saúde para predição diagnóstica de traços depressivos por meio de análise de dados em redes sociais.

Quanto à sua abordagem considera-se quantitativa, que segundo Marconi e Lakatos [1982] a define como a descrição objetiva sistemática e quantitativa do conteúdo manifesto da comunicação.

Para Marconi e Lakatos [1982], a análise quantitativa se efetua com toda informação numérica resultante da investigação", que se "apresentará como um conjunto de quadros, tabelas e medidas". Essa quantificação deve ser de forma sistemática e objetiva de acordo com, Marconi e Lakatos [1982]: "Sistemática: ser ordenada, metódica; Objetiva: proceder de forma rigorosa e reaplicável, ou seja, objetivação dos fenômenos, hierarquização das ações. Deve descrever, compreender e explicar os fatos".

Quanto aos objetivos compreende uma pesquisa exploratória e descritiva. Destaca-se o caráter de pesquisa exploratória com o objetivo de proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito como descreve Gil [2010]. Isto na medida em que demonstra que a partir de postagens e curtidas em redes sociais existe a possibilidade de explorar dados que denotam traços depressivos.

Classifica-se como descritiva, pois segundo Silva [2005], esta "visa descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis". Assim, tem característica descritiva, pois permitiu a obtenção de dados descritivos sobre pessoas, processos interativos" como nos explica Godoy [1995]. Onde a pesquisadora se ocupou diretamente em selecionar estas característica com a utilização da escola de Beck e apresentá-los como características de traços depressivos.

Quanto aos procedimentos caracteriza-se como bibliográfica, de levantamento e experimental. Trata-se de uma pesquisa bibliográfica na medida e que se utiliza de materiais como livros, revistas, artigos como explica Silva [2005]: "Pesquisa Bibliográfica: quando elaborada a partir de material já publicado, constituído principalmente de livros, artigos de periódicos e atualmente com material disponibilizado na Internet."

Corresponde à Pesquisa de levantamento na medida e que se utiliza de questionários semiestruturados, como explica Silva [2005]: "Levantamento: quando a pesquisa envolve a interrogação direta das pessoas cujo comportamento se deseja conhecer."

E por fim, caracteriza-se como pesquisa experimental, que segundo Wazlawick [2010] caracteriza-se pela manipulação de um aspecto da realidade pelo pesquisador, implicando em se ter uma ou mais variáveis experimentais que podem ser controladas pelo pesquisador, e uma ou mais variáveis observadas, cuja mensuração poderá levar, à conclusão de que existe algum tipo de dependência com a variável experimental. Sendo a pesquisa experimental adequada ao presente estudo por apresentar essa situação de manipulação, em que a variável experimental refere-se a dependente ou alvo representando os sintomas depressivos e as variáveis observadas referem-se àquelas relacionadas aos traços de sintomas e ao comportamento longitudinal dos usuários na rede social.

3.2 População e Amostra

A amostra selecionada para o estudo compreendeu usuários do Facebook com acesso ao Aplicativo Vivamente e os participantes foram acessados em dois períodos. A primeira etapa ocorreu entre maio a setembro de 2017 e a segunda etapa ocorreu entre outubro a dezembro desse mesmo ano. Para a seleção da amostra adotou-se os seguintes critérios de inclusão:

- Ser usuário voluntário com conta ativa na rede social Facebook;
- Ser residente e domiciliado em qualquer estado ou no Distrito Federal do Brasil;
- Com ou sem diagnóstico clínico depressivo;
- Ter idade acima de 18 anos;
- Aceitar os termos da pesquisa e condições de permissão de acesso aos dados;
- Com preenchimento do questionário e coleta de dados do usuário completa.

Em síntese, considerando ambas as etapas, houve 898 registros, com 692 (77%) foram considerados dentro do critério de inclusão. A amostra foi composta por 692 participantes (71.7% de mulheres e 28,3% de homens), com idade média de 29.1 anos (DP = 12.5), e que publicaram, em média, 1420 posts. Como os dados foram coletados desde 2011, foi possível verificar a frequência relativa de uso cada um dos anos, que foi de 2.1% em 2011, 5.5% em 2012, 12.6% em 2013, 15.3% em 2014, 18.8% em 2015, 23.6% em 2016 e 22.1 em 2017. Em relação aos meses do ano, em janeiro, a proporção de publicação foi de 7.8%, em fevereiro foi de 7.3%, em março foi de 8.1%, em abril foi de 8.3%, em maio foi de 9.4%, em junho foi de 8.5%, em julho foi de 8.8%, em agosto foi de 9.1%, em setembro foi de 8.9%, em outubro foi de 8.8%, em novembro foi de 8.1% e, finalmente, em dezembro foi de 6.9%. Aproximadamente, 10.2% das publicações foi feira pela manhã, 30.8% pela tarde, 24.5% pela noite e 33.5% pela madrugada. A Tabela 3.1 descreve as principais informações.

Sexo/P-valor	N (%)	Idade	Amigos(M|DP|)	Likes(M|DP|)	Posts(M|DP|)
Mulheres	496 (71.7%)	28.70 (12.90)	1026 (1060)	732 (932)	1638 (1409)
Homens	196 (28.3%)	30.30 (11.60)	974 (1130)	569 (871)	1521 (1449
P-valor	< 0.001	0.13	0.58	0.03	0.34

Tabela 3.1: *Dados sociodemográficos*
Fonte: Elaboração própria

Assim, a população envolvida na pesquisa compreendeu na primeira etapa, durante os meses de maio a setembro de 2017, dados de 473 usuários, sendo selecionados como amostra apenas 296 registros que atenderam a todos os critérios de inclusão. Na segunda etapa, de outubro a dezembro de 2017 foram realizadas mais 425 coletas, sendo destas incluídas na pesquisa 396. Totalizando 898 registros de coletas, onde 692 foram considerados válidos.

3.3 Métodos e técnicas aplicados à pesquisa

O método norteador da pesquisa refere-se ao experimental, em que segundo Rodrigues [2010], seu delineamento consiste em definir um objeto para estudo, selecionar as variáveis que seriam capazes de influenciá-lo, determinar as formas de controle e de observação dos efeitos que a variável escolhida produz no objeto, que na presente pesquisa o objeto refere-se aos sintomas depressivos e as variáveis que o influenciam são as que representam alguns traços depressivos e o comportamento do usuário na rede social.

Experimentos em computação assumem diversas formas, desde testes de desempenho de algoritmos à análise de fatores humanos, segundo Zobel [2004].

Na coleta de dados para a seleção dos participantes voluntários a amostragem aleatória simples foi empregada, resultando em uma amostra de uma população, bem como o método de levantamento (dados primários). Ainda foi aplicado questionário online autoaplicativo constituído por 21 questões fechadas com uma única opção de escolha. E, utilização do aplicativo Vivamente para acesso aos dados dos usuários em rede social. Já para a preparação, aplicação no modelo e análise dos dados foram utilizados:

- Escala de Depressão de Beck BDI-II descrita anteriormente neste estudo;
- Scripts MongoDB para limpeza e formatação dos dados, tornando-os adequados à utilização nos experimentos.

Durante o experimento do modelo proposto foi aplicado o Test T para analisar a diferença entre as médias de duas amostras, considerando uma amostra X1, X2, (com sintomas depressivos e sem sintomas depressivos). O coeficiente de correlação de Pearson para medir o grau e a direção da correlação de cada variável com relação a variável *label2*. A razão de Chance (RC), para calcular a razão entre a chance de um evento ocorrer em uma categoria de sintomas depressivos e a chance de ocorrer em outra categoria. O teste Wald foi aplicado para verificar a significância conjunta para cada variável estimada, identificando se cada uma das variáveis que se encontra no modelo explica, em termos estatísticos, os valores observados para a variável *label2*, com e sem sintomas depressivos, bem como para variável *label4* que representa as categorias minimo, leve, moderado e grave. Compreendendo assim a estatística descritiva utilizada para as análises preliminares dos dados e a estatística inferencial utilizada para as análises do modelo proposto.

3.4 Variáveis Selecionadas

Foram selecionadas 76 variáveis para a base de dados posts e para a base de dados likes, sendo 3 variáveis gerais (_id, sexo e idade), 4 variáveis derivadas das agregações, representando os totais (total_dias total_likes total_posts e total_friends), 4 variáveis derivadas das mensagens, item *message* (qtd_message, qtd_sent_neg e qtd_sent_pos), 1 variável derivada das ações do usuário, item *story* (qtd_story), 48 variáveis referentes ao tempo, item *created_time* (data, ano, dia, mes, hora, minutos, segundos, manha, tarde, noite, madrugada, t1, t2, t3, t4, dia_ano, dia_semana, segunda, terca, quarta, quinta, sexta, sabado, domingo, finalsemana, ano2011, ano2012, ano2013, ano2014, ano2015, ano2016, ano2017, janeiro, fevereiro, marco, abril, maio, junho, julho, agosto, setembro, outubro, novembro, dezembro e quantidadedevezesqueentrounodia), 21 variáveis referentes aos sintomas depressivos (agitacao, apetite, choro, concentracao, critica, culpa, desvalorizacao, energia, estima, fadiga, fracasso, indecisao, int_sexo, interesse, irritabilidade, pessimismo, prazer, punicao, sono, suicida e tristeza), 2 variáveis discretizadas derivadas da variável *nivel*, que representa o somatório dos níveis de depressão (label2 e label4).

As variáveis para a base de dados likes são muito semelhantes a base de dados posts, a única diferença é que nessa base de dados há uma variável derivadas das curtidas, item *likes* (qtd_likes) e as variáveis que representam as mensagens e ações do usuário não estão presentes nessa base de dados. Um apanhado completo de cada variável, tanto para a base de dados likes quanto a posts, com a sua respectiva descrição, encontram-se no Apêndice H.

Ressalta-se que dessas 76 variáveis selecionadas para cada uma das bases de dados, apenas 27 variáveis label4, prazer, fadiga, choro interesse desvalorizacao fracasso critica estima concentracao irritabilidade indecisao tristeza pessimismo suicida apetite sono culpa energia t1 t2 t4 total_friends total_likes total_posts sexo e idade) na base de dados posts, e somente 28 variáveis da base de dados likes (as 27 da base de dados posts e a variável madrugada) foram utilizadas na aplicação no modelo de regressão proposto. A ausência das variáveis no modelo, implica que, após os testes aplicados, as mesmas não convergiram, entrando em uma área de não concavidade. Para tanto foram utilizadas apenas 27 e 28 variáveis para as bases de dados posts e likes, respectivamente, as demais 49 e 48 foram descartadas.

3.5 Aprovação e autorização para realização da pesquisa

O projeto de pesquisa tramitou no Comitê de Ética em Pesquisa com Seres Humanos do Instituto de Psicologia da USP (CEPH-IPUSP)[1], enviado no dia 04 de janeiro de 2017 e aprovado, sob parecer de número 2.019.886, em 18 de Abril de 2017. O parecer do processo está detalhado no Apêndice A.

Antes da fase de desenvolvimento do *app* foram necessárias algumas autorizações. Para a utilização do BDI-II, por ser um instrumento utilizado por psicólogos, a Casa do Psicólogo, detentora dos direitos de publicação do BDI-II, autorizou a psicóloga Sabrina Zaffari Farias, CRP 12/10142, a responsabilizar-se pela aplicação online do referido instrumento como parte do método de pesquisa nesse trabalho de investigação científica.

- Projeto preenchido nos campos da própria Plataforma Brasil com o desenho, resumo, introdução, hipótese, objetivo, metodologia, riscos e desfecho.
- Currículo Lattes dos pesquisadores.

[1] http://www.ip.usp.br/

- Demonstrativo de existência de infraestrutura assinado pelo sr. diretor do Instituto de Matemática e Estatística (IME), Clodoaldo Grotta Ragazzo.
- Informações básicas do projeto, gerada pela plataforma.
- Folha de rosto gerada pela plataforma e assinada pelo sr. diretor do IME, Clodoaldo Grotta Ragazzo.
- Projeto de pesquisa detalhado, ou seja, além das partes já citadas no item 1, foram acrescentados os fundamentos, trabalhos relacionados, coleta de dados e cronograma.
- Termo de Consentimento Livre e Esclarecido (TCLE), contendo os termos relativos à pesquisa, tais como sua natureza, envolvimento, riscos, garantia de indenização, confidencialidade, benefícios, despesas e acompanhamento. O TCLE é assinado pelos pesquisadores e voluntários que aceitaram participar da pesquisa. O texto do TCLE está descrito na íntegra no Apêndice F.
- Declaração dos compromissos com os resultados da pesquisa.

3.6 Coleta de dados

Para a fase de coleta de dados para a pesquisa, o método selecionado foi o de acesso aos dados de uma rede social por meio de APIs, mais estritamente a GrapAPI do Facebook. Com o auxílio da API via Javascript um aplicativo (*app*) foi criado. Tal *app* precisou ser desenvolvido exclusivamente para essa fase da pesquisa pela necessidade em inserir um questionário inerente a investigação, bem como pela exigência do Facebook por permissões específicas para a extração de conteúdos de determinados itens.

O *app* Vivamente foi criado em 2 etapas, a primeira foi a configuração do ambiente no Facebook, e a segunda foi um projeto web desenvolvido e incorporado nesse ambiente.

O Facebook oferece várias funcionalidade de desenvolvimento à partir de suas APIs, como login pelo Facebook , plugins sociais, aplicativos embutidos, páginas externas, entre vários outros. Para a coleta de dados dessa pesquisa optou-se pelo desenvolvimento de um aplicativo a ser embutido no recurso Facebook *canvas*.

Na primeira etapa, relativa a configuração do ambiente no Facebook, algumas informações de conexão com o aplicativo web foram inseridas, como o nome do aplicativo, seu domínio, termos e política de privacidade, já a versão, o ID e a chave secreta (*app secret e app ID*) foram automaticamente fornecidas pelo Facebook.

Na segunda etapa, um projeto web foi criado, tendo sido produzido no ambiente *Sublime Text*. Foram também utilizados o *NodeJS*, como o interpretador de código JavaScript do lado do servidor; *Express*, como o framework para aplicativos *Node.js*; *Passport*, como *middleware* para autenticação da aplicação; *AngularJS*, para gerenciamento do questionário e *MongoDB* como banco de dados orientado a documentos. Nesse projeto, toda a lógica de programação foi desenvolvida, e através de alguns dos *scripts nodeJS* foram configuradas as autenticações e demais determinações para desenvolvimento de aplicativos sob a *Graph API* do Facebook. Uma cópia de todo o código fonte do *app* está armazenada no GitHub com o nome de Vivamente[2]. A arquitetura do aplicativo é mostrada de maneira sucinta na figura 3.1.

[2] https://github.com/cmaricy/Vivamente http://vivamente.herokuapp.com/

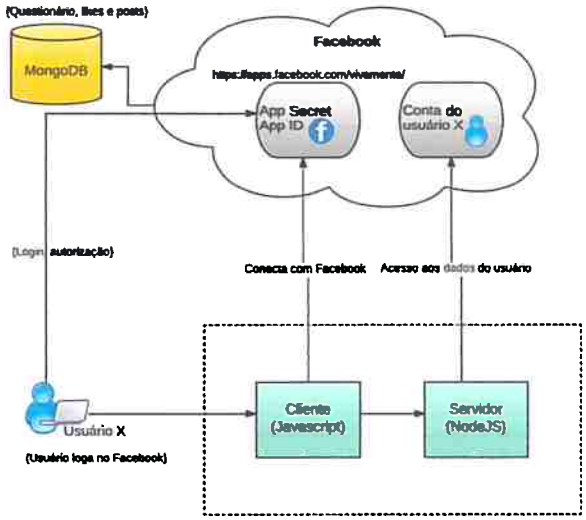


Figura 3.1: *Resumo da arquitetura do aplicativo Vivamente*
Fonte: Adaptado de Kzics

Finalizado o projeto, este foi hospedado em um servidor de aplicações web e seu *"Uniform Resource Locator"* (URL) foi configurado dentro do *Facebook canvas* e passou a ser executado à partir desse ambiente sob o URL[3] fornecido pelo Facebook, concluindo, portanto, o *app* de coleta denominado Vivamente. Os detalhes do processo de desenvolvimento do *app* estão descritos no Apêndice B.2.

Posteriormente ao término do *app*, alguns testes foram realizados para verificar sua consistência, testado e executando corretamente, foi preciso submetê-lo a revisão do Facebook, pois a coleta dos dados dos itens *user_likes* e *user_posts* não é autorizada por padrão, sendo necessária a aprovação específica para esses itens, detalhes do processo de solicitação e aprovação estão no C.

Estando o app Vivamente sendo executado de maneira precisa, iniciou o processo de divulgação junto aos usuários finais, a fim de obter voluntários dispostos a participar da pesquisa.

A divulgação para alcançar os voluntários da pesquisa foi realizada primeiramente por meio de divulgações via e-mail, por postagem e compartilhamento pela linha do tempo no Facebook, via Linkedin e Twitter. O texto convidando a participar finalizava com o link para a pesquisa. Essa forma de divulgação foi realizada durante os meses de maio e junho, dois meses portanto, e resultou em 202 instalações do app vivamente, ou seja, 202 registros de usuário.

Em um segundo momento foi criado um vídeo no Youtube intitulado "A depressão pode levar ao suicídio!"e divulgado na linha do tempo do Facebook, essa ação teve um retorno de 10 registros de usuário.

Para o período final de coleta uma página no Facebook foi criada sob a categoria de comunidade, tendo sido promovida pelo Facebook pelo período de 10 dias, alcançou 181 pessoas a um custo de 8,00 reais e obteve 7 registros de usuário. Na sequência o próprio *app* vivamente foi promovido em uma campanha do Facebook alcançando um público de 14.616 pessoas, a um custo de 145,51 reais onde obteve 114 instalações do aplicativo e por consequência 114 registros de usuário. Finalmente o vídeo do Youtube foi também promovido por um período de 21 dias, alcançando 2.369 visualizações, a um custo de 100,00 reais e obteve 5 registros de usuário. Mais detalhes da trajetória da divulgação do app encontram-se no Apêndice D.

A campanha em um primeiro momento alcançou 296 usuários, um número abaixo das expectativas e além do mais a base de dados estava bastante desbalanceada no sentido de haver muitos usuários sem sintomas depressivos e poucos usuários apresentando os sintomas, sendo assim o período de coleta se extendeu por mais 4 meses, onde além dos meios de divulgação anterior, campanhas no Facebook e em comentários de postagens relacionadas ao assunto "depressão", tendo alcançado 898 registros de usuário, sendo 692 registros válidos, embora novamente desbalanceado, dessa vez com mais usuário com sintomas depressivos do que o contrário, mas o número foi considerado adequado para a exploração.

3.6.1 Conjunto de dados Vivamente

Os dados obtidos na fase de coleta formaram um conjunto ou base de dados, também chamado de *dataset*, de acordo com van der [2016].

[3] https://apps.facebook.com/vivamente

O *dataset* obtido é um conjunto de registros de usuário (também conhecido como objeto, arquivo, caso, linha, entre outros, dependendo da área ou do domínio de estudo). Neste domínio, os dados são provenientes do Facebook, e aqui referido como *dataset Vivamente*, formado por uma coleção de registros de usuário, às vezes denominado registro ou simplesmente usuário, onde cada registro refere-se aos dados de um usuário. Cada registro contém 6 atributos independentes: _id, created_time, autoriza, idade, sexo, nome, friends e id_usuario, esses atributos recebem apenas um valor cada. Possui também 3 arrays de objetos, um chamado respostas com 22 atributos sendo: _id, created_time, agitacao, apetite, choro, concentracao, critica, culpa, desvalorizacao, energia, estima, int_sexo, fadiga, fracasso, indecisao, intSexo, interesse, irritabilidade, pessimismo, prazer, punicao, sono, suicida, tristeza e nivel, atributos que também recebem apenas um valor. Outra coleção chamada posts com 5 atributos: _id, id, created_time, message e story, esses atributos formam um objeto, e a coleção de objetos pode variar em seu total pois depende da quantidade de postagens do usuário, o atributo message ou story podem não estar presentes em todos os objetos. A última coleção chamada likes contém 4 atributos: _id, id, created_time e name, também formam objetos e todos os atributos estão presentes em todos os objetos. A figura 3.2 é uma amostra dos dados com visualização em formato de tabela, onde pode-se observar que alguns atributos são independentes e outros estão em formato de arrays de objetos.

_id	created_time	autoriza	idade	sexo	id_usuario	respostas	posts	likes	friends	public_profile
5a01b...	2017-1...	S	48	F	7690...	[1 el...	[1397 e...	[2211...	{ 2 fiel...	{ 13 fiel
5a024...	2017-1...	S	56	M	1549...	[1 el...	[1057 e...	[285 el...	{ 2 fiel...	{ 13 fiel
5a027...	2017-1...	S	42	F	8725...	[1 el...	[2214 e...	[258 el...	{ 2 fiel...	{ 13 fiel
5a030...	2017-1...	S	27	F	1595...	[1 el...	[5295 e...	[1154...	{ 2 fiel...	{ 13 fiel
5a035...	2017-1...	S	19	M	1537...	[1 el...	[3324 e...	[1392...	{ 2 fiel...	{ 13 fiel
5a035...	2017-1...	S	43	F	5181...	[1 el...	[1757 e...	[427 el...	{ 2 fiel...	{ 13 fiel
5a035...	2017-1...	S	34	F	9036...	[1 el...	[2391 e...	[275 el...	{ 2 fiel...	{ 13 fiel
5a035...	2017-1...	S	31	F	9672...	[1 el...	[2823 e...	[972 el...	{ 2 fiel...	{ 13 fiel
5a037...	2017-1...	S	48	F	1675...	[1 el...	[3730 e...	[252 el...	{ 2 fiel...	{ 13 fiel
5a039...	2017-1...	S	25	F	1498...	[1 el...	[1628 e...	[499 el...	{ 2 fiel...	{ 13 fiel
5a039...	2017-1...	S	32	F	1545...	[1 el...	[812 ele...	[698 el...	{ 2 fiel...	{ 13 fiel
5a03a...	2017-1...	S	52	F	1587...	[1 el...	[748 ele...	[546 el...	{ 2 fiel...	{ 13 fiel
5a03a...	2017-1...	S	18	F	8699...	[1 el...	[903 ele...	[457 el...	{ 2 fiel...	{ 13 fiel
5a03b...	2017-1...	S	30	F	1674...	[1 el...	[3875 e...	[1251...	{ 2 fiel...	{ 13 fiel
5a03c...	2017-1...	S	48	F	1493...	[1 el...	[1519 e...	[183 el...	{ 2 fiel...	{ 13 fiel
5a03d...	2017-1...	S	23	F	1225...	[1 el...	[2913 e...	[599 el...	{ 2 fiel...	{ 13 fiel
5a043...	2017-1...	S	20	F	1178...	[1 el...	[977 ele...	[563 el...	{ 2 fiel...	{ 13 fiel
5a043...	2017-1...	S	21	F	1492...	[1 el...	[1901 e...	[272 el...	{ 2 fiel...	{ 13 fiel
5a044...	2017-1...	S	19	F	1701...	[1 el...	[4263 e...	[1337...	{ 2 fiel...	{ 13 fiel
5a045...	2017-1...	S	21	F	1566...	[1 el...	[705 ele...	[167 el...	{ 2 fiel...	{ 13 fiel
5a045...	2017-1...	S	47	F	1527...	[1 el...	[5322 e...	[837 el...	{ 2 fiel...	{ 13 fiel
5a046...	2017-1...	S	23	F	3872...	[1 el...	[90 ele...	[740 el...	{ 2 fiel...	{ 13 fiel
5a046...	2017-1...	S	31	F	1432...	[1 el...	[2954 e...	[644 el...	{ 2 fiel...	{ 13 fiel

Figura 3.2: *Amostra dos dados com visualização em formato de tabela*
Fonte: Elaboração própria

A figura 3.3 mostra um registro dos dados com visualização em formato de árvore, onde os objetos estão denominados como *elements* dentro dos arrays.

Key	Value	Type
▼ (1) {_id : 590bc2df23...	{ 13 fields }	Document
_id	590bc2df239ac8004300d283	ObjectId
created_time	2017-05-05T00:10:07.700Z	Date
autoriza	S	String
idade	46	Int32
sexo	F	String
nome	Maricy Caregnato	String
id_usuario	1104080929631843	String
▶ respostas	[1 elements]	Array
▶ posts	[1745 elements]	Array
▶ likes	[112 elements]	Array
▶ friends	{ 2 fields }	Object
▶ public_profile	{ 13 fields }	Object

Figura 3.3: *Registro dos dados com visualização em árvore*
Fonte: Elaboração própria

Os dados, podem ser armazenados em diferentes formatos, como tabelas, planilhas, arquivos, entre outros. O *dataset Vivamente* original foi armazenado em um banco de dados MongoDB, que segundo Fowler

[2012], é um banco de dados de documentos NoSQL com uma série de recursos que no mundo do software livre (*open source*), são difíceis de serem superados. O banco de dados MongoDB manuseia nativamente documentos JSON, sendo o formato que o Facebook fornece os seus dados.

A tabela 3.2 mostra trechos dos dados de um registro extraído do Facebook em formato JSON, onde pode-se visualizar primeiramente à esquerda os atributos independentes, à sua direita, o array respostas com alguns dos 21 atributos relativos aos sintomas da depressão, abaixo uma amostra do array referente às postagens (posts) com dois objetos (com seus atributos e valores), e à sua direita o array referente às curtidas (likes) seguindo o mesmo padrão.


```
{
   "_id" : ObjectId("590bc2df239ac8004300d283"),
   "created_time" : ISODate("2017-05-05T00:10:07.700+0000"),
   "autoriza" : "S",
   "idade" : NumberInt(46),
   "sexo" : "F",
   "nome" : "Fulana",
   "id_usuario" : "1104080929631843",
```

```
"respostas" : [
    {
       "created_time" : ISODate("2017-05-05T00:10:07.706+0000"),
       "nivel" : NumberInt(8),
       "tristeza" : "0",
       "pessimismo" : "0",
       ...
       "suicida" : "0",
       "agitacao" : "1",
       "interesse" : "0",
       "energia" : "1",
       "sono" : "1",
       ...
       "concentracao" : "1",
       "fadiga" : "1",
       "_id" : ObjectId("590bc2df239ac8004300d284")
    }
          ],
```

```
"posts" : [  // array de objetos posts
   {  // inicia um objeto
      "message" : "Sem mais...",  // atributo
      "created_time" : ISODate("2017-04-21T03:00:22.000+0000"),
      "id" : "1104080929631843_1492837554089510",
      "_id" : ObjectId("590bc2f6239ac8004300d285")
   },  // termina um objeto
   {
      "message" : "Mudança de comportamento pode ser a chave!",
      "story" : "Beltrano  added 3 new photos — with Fulana .",
      "created_time" : ISODate("2017-04-20T16:01:10.000+0000"),
      "id" : "1104080929631843_1492281070811825",
      "_id" : ObjectId("590bc2f6239ac8004300d286")
   },
```

```
"likes" : [
   {
      "name" : "Condominio Paço da Universidade",
      "id" : "272810919416244",
      "created_time" : ISODate("2017-04-15T20:03:27.000+0000"),
      "_id" : ObjectId("590bc2f6239ac8004300d958")
   },
   {
      "name" : "Journal of Web Semantics",
      "id" : "181730910961",
      "created_time" : ISODate("2016-07-28T14:46:02.000+0000"),
      "_id" : ObjectId("590bc2f6239ac8004300d95f")
   },
```


Tabela 3.2: *Trechos dos dados de um registro em formato JSON.*
Fonte: Elaboração própria

Segundo Cielen e Meysman [2016], os dados obtidos na fase de coleta de dados geralmente estão em uma forma bruta, sendo preciso limpá-los e prepará-los para uso nas fases de análise e apresentação, sendo um bom hábito corrigir os erros de dados tão cedo quanto possível no processo.

Algumas medidas de precaução foram tomadas na fase de desenvolvimento do *app* de coleta Vivamente, dentre elas a impossibilidade de deixar qualquer um dos campos sem preenchimento ou a aceitação de valores diferentes do definido, a idade por exemplo, só pode variar de 18 a 100, e registros com _ids duplicados foram rejeitados.

Outras ações precisaram ser efetuadas para fornecer a garantia da usabilidade dos dados no projeto, o que significou realizar uma etapa de preparação, visando a limpeza e afiltragem dos dados.

3.7 Preparação dos dados

Na fase da preparação dos dados deverá haver um aumento na qualidade dos dados e sua preparação para uso em etapas subseqüentes, conforme define Cielen e Meysman [2016].

Para garantir que os dados estivessem em um formato adequado para uso no modelo proposto a primeira providência foi a busca e exclusão de registros com arrays vazios, esse suposto erro ocorreu intencionalmente (programado) porque se caso houvesse alguma falha na comunicação ou cancelamento da varredura dos dados da linha do tempo do usuário durante a coleta, os arrays zerariam, como forma de evitar dados com quantidades de atributos equivocadas.

A figura 3.4 mostra alguns registros em que os arrays likes e posts estão vazios.

_id	created_time	autoriza	idade	sexo	id_usuario	respostas	posts	likes
5928...	2017-...	S	37	F	102...	[1 el...	[0 elements]	[0 elements]
592c...	2017-...	S	26	M	102...	[1 el...	[0 elements]	[0 elements]
5930...	2017-...	S	22	M	138...	[1 el...	[0 elements]	[0 elements]
5935...	2017-...	S	32	M	101...	[1 el...	[0 elements]	[0 elements]
5943...	2017-...	S	39	F	454...	[1 el...	[0 elements]	[0 elements]
59c2...	2017-...	S	38	F	152...	[1 el...	[0 elements]	[0 elements]
59c2...	2017-...	S	26	F	135...	[1 el...	[0 elements]	[0 elements]
59c3...	2017-...	S	37	F	177...	[1 el...	[0 elements]	[0 elements]
5a06...	2017-...	S	21	F	199...	[1 el...	[0 elements]	[0 elements]
5a07...	2017-...	S	30	F	145...	[1 el...	[0 elements]	[0 elements]

Figura 3.4: *Registros com arrays vazios*
Fonte: Elaboração própria

Para essa etapa de busca e exclusão foram criadas *queries* específicas direto no *dataset Vivamente*, pela sua simplicidade e leveza, conforme mostra a figura 3.5.

```
1
2 db.vivamente.aggregate([
3
4 {$match:{'$posts':{$gt:{} }}}
5
6 ])
7
```

Figura 3.5: *Query para exclusão de documentos com array vazio*
Fonte: Elaboração própria

Em seguida foi realizada uma exclusão de atributos que em nada acrescentariam às análises (a exemplo dos atributos nome, public_profile e ids dos objetos internos) ou que não poderiam estar visíveis (como foi o caso do atributo nome).

Foram também executados vários comandos para filtragem dos dados, a exemplo da agregação que busca alguns atributos e soma outros, criando um primeiro *dataset* nomeado *Vivamente Agregado* a ser utilizado posteriormente. A figura 3.6 mostra um pedaço da *query* e do seu resultado.

```
db.vivamente.aggregate([{
    $project: {
        idade: "$idade",
        sexo: "$sexo",
        qtd_friends: "$friends.summary.total_count",
        totalPosts :{$size : "$posts"},
        totalLikes :{$size : "$likes"} ,
        dataIni: { $dateToString: { format: "%d-%m-%Y%H:%M:%S", date: {$min:"$posts.created_time" }}},
        dataFim: { $dateToString: { format: "%d-%m-%Y%H:%M:%S", date: {$max: "$posts.created_time" }}},
        dias: {$divide: [{"$subtract": [{"$max":"$posts.created_time"}, {"$min": "$posts.created_time"}] } , 86400000]},
        nivel: { $arrayElemAt: ["$respostas.nivel", 0]},
        agitacao: { $arrayElemAt: ["$respostas.agitacao", 0]},
        apetite: { $arrayElemAt: ["$respostas.apetite", 0]},
        choro: { $arrayElemAt: ["$respostas.choro", 0]},
```

Aggregate | Aggregate

296 | Documents 1 to 296

Root Level

_id	idade	sexo	qtd_friends	totalPosts	totalLikes	dataIni	dataFim	dias	nivel	agitacao	apetite	choro
59...	39	M	293	2738	54	01-01-20060...	16-06-2017...	4184....	1	0	0	0
59...	40	M	548	3058	446	01-01-2009...	07-05-2017...	3048....	5	0	0	0
5a...	19	M	3877	3324	1392	01-01-2010...	08-11-2017...	2868....	37	3	2	3
5a...	19	F	863	2427	1493	01-01-2010...	11-11-2017...	2870....	47	3	3	2
59...	32	F	687	3792	580	01-01-2010...	29-05-2017...	2705....	13	0	0	0
59...	32	F	946	1951	534	01-01-2011...	13-09-2017...	2447....	7	0	0	0
5a...	18	F	4928	1476	2041	01-01-2012...	10-11-2017...	2139....	51	1	3	2
59...	30	M	1896	939	188	01-01-2012...	18-06-2017...	1995....	47	3	2	3
5a...	18	F	1097	903	457	01-01-2013...	09-11-2017...	1772....	43	3	2	3
5a...	23	F	432	457	759	01-01-2014...	09-11-2017...	1408....	47	2	2	3
59...	24	F	752	100	297	01-01-2017...	08-05-2017...	127.2...	0	0	0	0
5a...	20	F	3403	608	1148	01-04-2012...	10-11-2017...	2049....	45	3	3	2
5a...	21	F	545	1901	272	01-05-2011...	08-11-2017...	2383....	45	2	2	3
5a...	34	F	233	2391	275	01-09-2014...	08-11-2017...	1164....	29	1	2	1

Figura 3.6: *Query para busca de documentos com atributos específicos*
Fonte: Elaboração própria

Aqui foi criada uma *query* para agregar o total de ações do usuário por categoria, que estão armazenadas no atributo story do array posts, como por exemplo número de vezes que o usuário trocou de foto de capa, de foto de perfil, postou fotos com amigos, mudou de relacionamento, compartilhou vídeos, estava sentindo-se feliz, triste, orgulhoso, entre outras ações. Uma amostra dessa *query* e seu resultado pode ser visualizado na figura 3.7.

```
var cursor = db.vivamente.find();
while (cursor.hasNext()){
  var doc = cursor.next(),
            qtd_profile = 0,
            qtd_cover = 0,
            qtd_shared_photo = 0,
            qtd_shared_link = 0;
  doc.posts.forEach(function(post){
    if (post.story){
        if (post.story.toLowerCase().indexOf('updated his profile picture') > -1
          || post.story.toLowerCase().indexOf('updated her profile picture') > -1 ) qtd_profile++;

      if (post.story.toLowerCase().indexOf('updated his cover photo') > -1
          || post.story.toLowerCase().indexOf('updated her cover photo') > -1) qtd_cover++;
  var obj = {};
```

Text | Documents

50 | Documents 1 to 50

Root Level

_id	qtd_update_prof	qtd_update_cover	qtd_shared_photo	qtd_shared_link	qtd_message	qtd_story	qtd_add_photo
590bc...	35.0	20.0	182.0	17.0	1445.0	876.0	323.0
590fa...	23.0	10.0	450.0	39.0	1505.0	927.0	150.0
590fd...	11.0	4.0	722.0	120.0	2397.0	1390.0	75.0
5911d...	4.0	1.0	15.0	0.0	78.0	85.0	12.0
59125...	9.0	3.0	737.0	378.0	147.0	1462.0	85.0
59220...	9.0	13.0	29.0	2.0	680.0	344.0	63.0
5923a...	16.0	4.0	247.0	46.0	3054.0	1442.0	551.0
59242...	0.0	1.0	21.0	39.0	364.0	134.0	20.0
59242...	3.0	1.0	1.0	1.0	79.0	60.0	10.0
59243...	2.0	1.0	2.0	2.0	47.0	61.0	27.0
59243...	13.0	18.0	282.0	30.0	1225.0	1010.0	123.0
59243...	9.0	9.0	70.0	101.0	1384.0	346.0	47.0

Figura 3.7: *Uma query para agrupar as ações do usuário*
Fonte: Elaboração própria

O próximo exemplo mostra uma *query* que desmembra o atributo created_time do array posts em ano, mês, dia, hora, minutos e segundos, e faz uma contagem de cada post para uma referida data, a fim de transformar essas variáveis em um padrão de dados longitudinais. A figura 3.8 mostra um trecho da *query* e do seu resultado.

```
db.vivamente.aggregate([
    { $match : { 'posts.created_time' :
     { $gte: ISODate("2005-01-01T00:00:00.000Z") } } },
     {$unwind: '$posts' },
    { $match : { 'posts.created_time' :
     { $gte: ISODate("2005-01-01T00:00:00.000Z") } } },
    { $group : {
      _id : {
                  year : { $year : '$posts.created_time' },
                  month : { $month : '$posts.created_time' },
                  day : { $dayOfMonth : '$posts.created_time' },
                  hour : { $hour : '$posts.created_time' },
                  minutes: { $minute: "$posts.created_time" },
                  seconds: { $second: "$posts.created_time" },
                  dayOfYear: { $dayOfYear: "$posts.created_time" },
                  dayOfWeek: { $dayOfWeek: "$posts.created_time" },
                 _id: '$_id' },
   total : { $sum : 1 }
```

Aggregate

50 | Documents 1 to 50

```
{
    "_id" : {
        "year" : NumberInt(2006),
        "month" : NumberInt(1),
        "day" : NumberInt(1),
        "hour" : NumberInt(8),
        "minutes" : NumberInt(0),
        "seconds" : NumberInt(0),
        "dayOfYear" : NumberInt(1),
        "dayOfWeek" : NumberInt(1),
        "_id" : ObjectId("5947e0f57ef102001db94b2e")
    },
    "total" : 3.0
}
{
    "_id" : {
        "year" : NumberInt(2007),
        "month" : NumberInt(1),
        "day" : NumberInt(1),
```

Figura 3.8: *Uma query para desmembrar o atributo created_time*
Fonte: Elaboração própria

O exemplo na figura 3.8 , mostra um trecho da *query* utilizada na base de dados *Vivamente*. Vários outros tipos de agregações foram realizados no *dataset Vivamente*, aqui foram mostrados apenas alguns para exemplificar a forma como foi trabalhada essa etapa.

3.8 Exploração dos dados

Schutt *et al.* [2016], define a exploração de dados como o início para construir um modelo. O aspecto "exploratório"significa que a compreensão do problema que se está resolvendo, ou pode resolver, está mudando à medida que avança.

Cielen e Meysman [2016], contextualiza que durante a Exploração de Dados (EDA), há um aprofundamento nos dados.

3.8.1 Exploração dos dados das postagens

Após algumas transformações, agregações e desmembramentos no *dataset Vivamente*, a etapa subsequente foi a criação de uma *query* que filtrou o *dataset* e retornou os textos (atributo message dos objetos do array posts) com suas respectivas datas em vários formatos e subdivisões, transformando-a em uma nova coleção, porém ainda em formato JSON, o próximo passo foi transformar essa nova base em um formato adequado a ser consumido por uma ferramenta de mineração de dados. A figura 3.9 contém uma amostra do formato da nova base.

ExampleSet (406805 examples, 0 special attributes, 11 regular attributes)

_id	day	mo...	year	hour	min...	sec...	day...	day...	text	total
592...	26	5	2009	14	32	28	3	146	uma fejuca ...	1
592...	26	5	2009	19	54	18	3	146	sinto orgulh...	1
592...	28	5	2009	18	12	38	5	148	todo cuidad...	1
592...	28	5	2009	18	56	3	5	148	histórias me...	1
592...	4	6	2009	20	20	33	5	155	prodígio? ou...	1
592...	4	6	2009	22	1	31	5	155	Eu e os ami...	1
592...	5	6	2009	14	7	31	6	156	frio no ranc...	1
593...	30	5	2009	1	3	50	7	150	As grandes i...	1
592...	8	6	2009	20	31	26	2	159	creme de m...	1
594...	9	6	2009	19	35	52	3	160	Hi how are y...	1

Figura 3.9: *Base de dados com textos e datas*
Fonte: Elaboração própria

Após a extração dos textos e suas respectivas datas, foi realizada uma classificação da polaridade dos sentimentos dos textos referentes a de cada uma das postagens, para isso, alguns testes com diferentes algoritmos de classificação de textos, como o algoritmo de *Deep Learning* usando H2O 3.8.2.6, o algoritmo ID3 de árvore de decisão, o *Naive Bayes* e o *SVM* de Stefan Rueping, foram realizados. Algumas variações de configurações foram testadas em cada um dos algoritmos, porém a maior diferença na confiabilidade foi relacionada a etapa de pré processamento de texto.

Os primeiros testes foram realizados com a utilização do dicionário léxico WordNet 3.0, com as etapas de *tokenize, lower case, filter stopwords* e *steeming* e para a extração dos sentimentos foi utilizado o *SentiWordNet 3.0*, conforme mostra a figura 3.10.

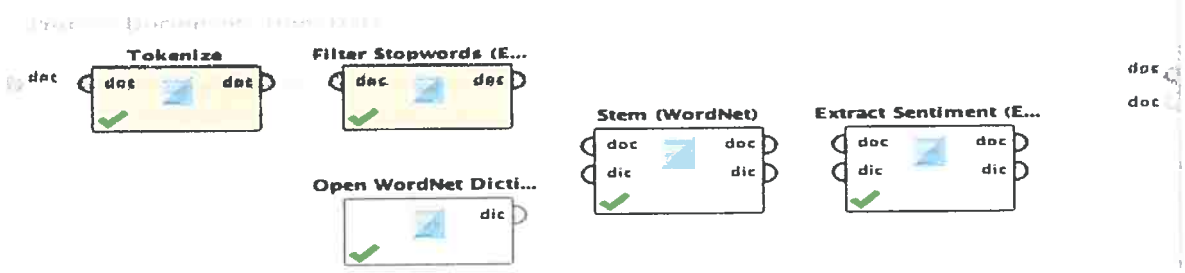


Figura 3.10: *WordNet e SentiWordNet para polarização dos sentimentos*
Fonte: Elaboração própria

A avaliação dos algoritmos de aprendizagem foi realizada por meio da análise de desempenho utilizando Validação Cruzada (***cross validation***), conforme pode ser visto no processo de classificação na figura 3.11 .

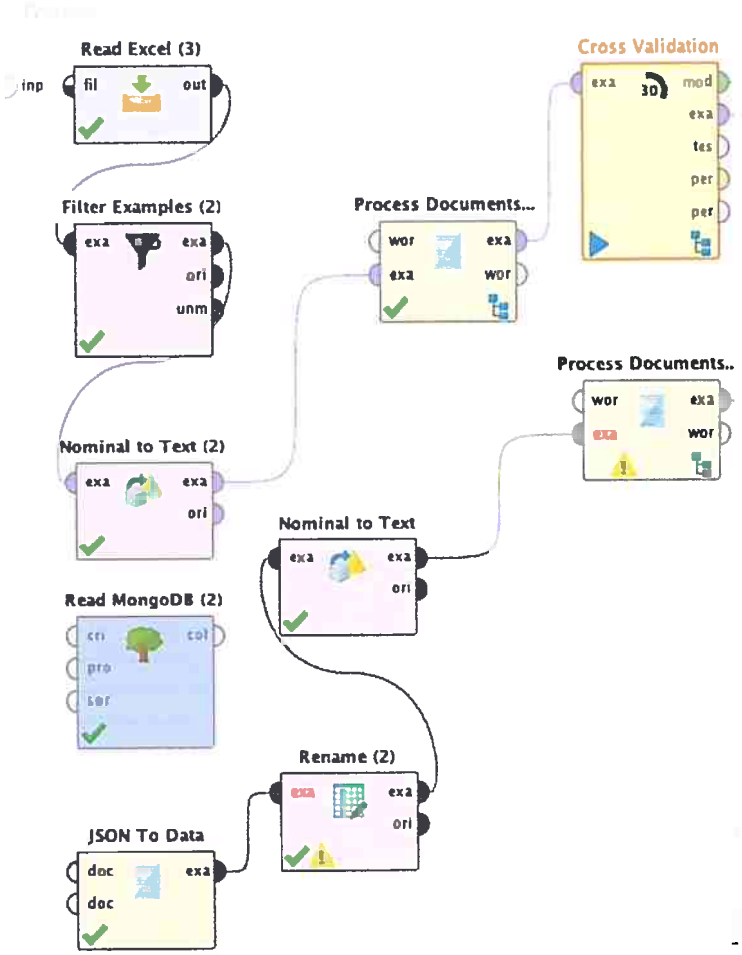


Figura 3.11: *Processo de avaliação do classificador utilizando validação cruzada*
Fonte: Elaboração própria

Com a utilização do *dataset WordNet*, obteve-se uma acurácia e precisão de 79,33 e 80% respectivamente, conforme mostram as figuras 3.12 e 3.13.

accuracy: 79.33%

	true positivos	true negativos	class precision
pred. positivos	240	65	78.69%
pred. negativos	59	236	80.00%
class recall	80.27%	78.41%	

Figura 3.12: *Acurácia do classificador SVM com WordNet.*
Fonte: Elaboração própria

precision: 80.00% (positive class: negativos)

	true positivos	true negativos	class precision
pred. positivos	240	65	78.69%
pred. negativos	59	236	80.00%
class recall	80.27%	78.41%	

Figura 3.13: *Precisão do classificador SVM com WordNet.*
Fonte: Elaboração própria

Como a precisão e a acurácia não estavam satisfatórias, outras alternativas precisaram ser testadas para tentar melhorar esses índices. A primeira decisão foi testar a classificação com uma base de dados de treino específica, até então estava sendo utilizada uma base genérica, em inglês do próprio *WordNet*.

Iniciando pela troca do idioma, duas base de dados em português do Brasil rotuladas com dados oriundos do Twitter foram testadas, uma delas para o domínio político e a outra para classificação de opiniões sobre filmes, tendo sido obtidos acréscimos em média de 4% para a precisão e 5% para a acurácia, indicando que poderia ser um caminho a ser explorado. Sendo assim, a próxima decisão foi a criação de uma base de dados rotulada com sentimentos positivos e negativos com os textos dos próprios atributos *message* de cada objeto do *array posts*.

A base de dados foi rotulada manualmente utilizando a plataforma crowdflower[4] tendo sido gerada uma base de treino rotulada (sentimentos positivos e negativos) com 10.000 linhas. Essa base foi submetida à avaliação, primeiramente da psicóloga responsável pelo projeto e posteriormente por outro psicólogo participante e por fim por um acadêmico do curso de psicologia, colaborador do projeto auxiliado pela sua orientadora na Universidade Federal de Mato Grosso (UFMT). O *overfitting*, que geralmente é uma preocupação no sentido de que os dados poderiam estar superajustados, não foi levada em consideração nessa fase pois os textos foram classificados para acrescentar mais informações ao *dataset Posts* (nova base de dados desmembrada do *Vivamente*, sendo gerada pelas postagens e desvinculada dos dados das curtidas), que deverá ser utilizada nas próximas etapas deste trabalho.

Em uma nova série de testes, utilizando dessa vez o *dataset Posts*, os resultados foram satisfatórios, chegando a uma precisão de 94,37% , com a utilização do classificador SVM, conforme mostra a figura 3.14.

precision: 94.37% +/- 1.10% (mikro: 94.35%) (positive class: negativo)

	true positivo	true negativo	class precision
pred. positivo	839	161	83.90%
pred. negativo	152	2538	94.35%
class recall	84.66%	94.03%	

Figura 3.14: *Precisão do classificador SVM e base de treino específica*
Fonte: Elaboração própria

O classificador H2O utilizando *Deep Learning* obteve uma acurácia e precisão superiores aos do SVM em média 2% , chegando a 96,4, porém a morosidade acabou inviabilizando a continuidade dos testes e sua possível adoção.

Ao final dessa etapa, obteve-se um *datasets* em forma de tabela, chamado *Sentimentos/Message* com 13 atributos desmembrados em datas (ano, mês, dia, hora, minutos, segundos, dia do ano, etc) referentes ao atributo *message* e também com os textos das mensagens classificados em positivos e negativos, representados por um atributo chamado ***prediction***, bem como seus níveis de confidencialidade, tanto para os sentimentos positivos quanto para os negativos, representados pelos atributos *confidencialidade(positivo)* e *confidencialidade(negativo)*, respectivamente, conforme podem ser visualizados na figura 3.15.

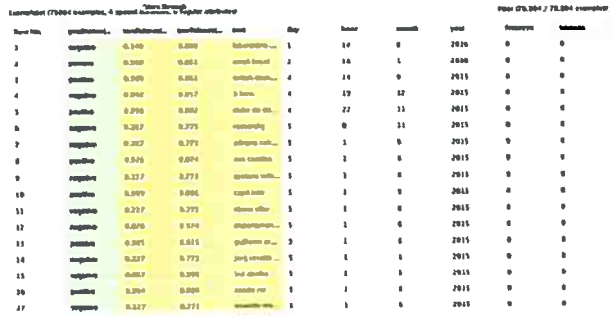


Figura 3.15: *Sentimentos classificados com SVM e base de treino específica*
Fonte: Elaboração própria

Após essa etapa da extração da polaridade dos sentimentos, outra parte da base de dados *Vivamente* precisou ser trabalhada, com os dados do atributo *story*, ou seja, dos atributos referentes as ações dos usuá-

[4] https://www.crowdflower.com

rios, por ocorrerem em datas diferentes do atributo *message*, como nessa situação não houve a necessidade de extração de outras informações relacionadas as ações, essa base de dados simplesmente foi unificada à base de dados *Sentimentos/Message*, formando a base de dados referente às postagens *(story e message)* com os sentimentos (que referem-se ao atributo *message* e os demais atributos, essa nova base é referida como base de dados *Posts* e pode ser visualizada na figura 3.16.

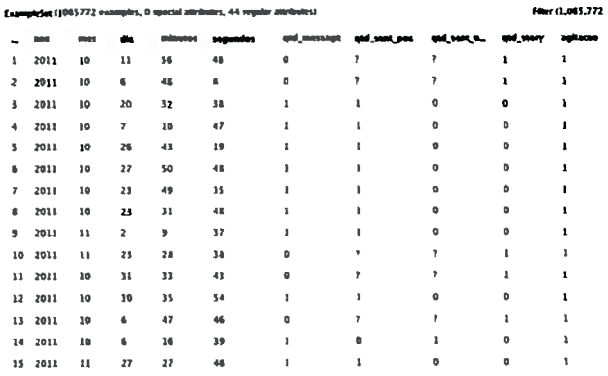


Figura 3.16: *Base de dados longitudinal referente às postagens*
Fonte: Elaboração própria

Observa-se, por meio da figura 3.16, que a base de dados *Posts* possui agora 1.085,772 registros, transformando-se em uma base de dados longitudinal, onde são considerados os dados temporais de cada usuário.

A próxima etapa de transformação nos dados foi a retirada de *missings*, (que ocorreram pela entrada dos dados do cálculo do atributo *story* nomeado de *qtd story*). Na sequência duas transformação foram realizadas, a primeira foi uma discretizações referente ao atributo *nivel*, tendo sido classificado em 4 categorias de sintomas depressivos (mínimo, leve, moderado e grave) representados pelo atributo *label4*, e posteriormente a binarização do cálculo dos sete atributos considerados independentes da função física, gerando assim o atributo *label2*, o resultado pode ser visualizado na figura 3.17.

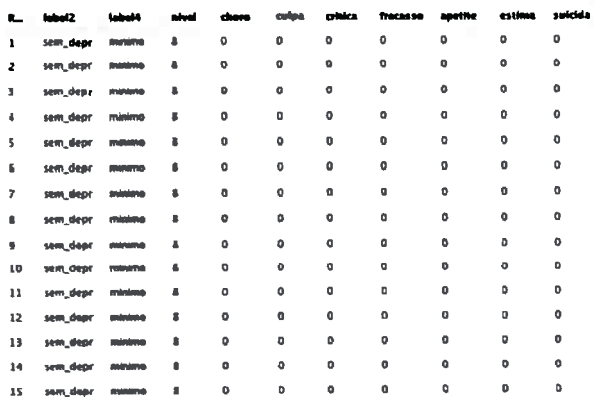


Figura 3.17: *Base de dados Posts discretizada*
Fonte: Elaboração própria

A última etapa que precisou ser aplicada aos dados foi uma agregação temporal, tornando algumas variávies *dummies*, para que os dados pudessem estar bem adequados à aplicação dos modelos de teste na fase da análise (processamento ou mineração dos dados).

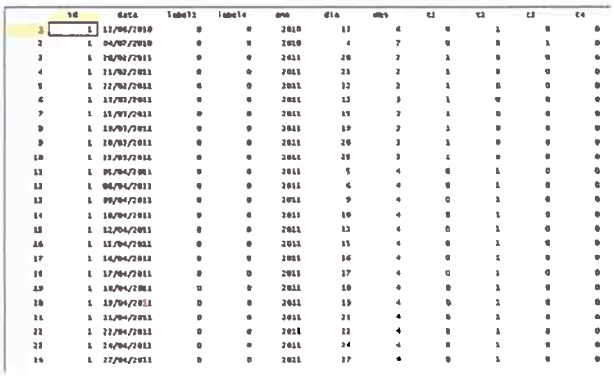


Figura 3.18: *Base de dados Posts temporal dummy*
Fonte: Elaboração própria

A figura 3.18, mostra apenas uma pequena parte, com exemplo dos atributos trimestrais (t1, t2, t3 e t4), porém vários outros atributos sofreram esse processo e alguns nomes de atributos foram alterados a fim de torná-los mais intuitivos, a relação completa dos atributos, bem como a descrição de cada um encontra-se no Apêndice H. Seguindo essas mesma utilizadas para a criação do *dataset Posts*, o outro *dataset* chamado *Likes* foi criado, dessa vez para os dados relacionados às curtidas contidas no *array* de dados *likes*. Mais detalhes sobre esses *datasets*, bem como as bases de dados completas encontram-se no armazenados no *github*, no repositório sob o nome de *datasets Vivamente*.

Capítulo 4

Análises, resultados e discussões

Este capítulo demonstra as etapas desta investigação cujo objeto de interesse consiste em apresentar um Modelo de Regressão Logística Multinomial, considerando o comportamento do usuário em rede social, para a predição de probabilidades de traços depressivos.

Seguindo a teoria referente aos modelos logísticos exposta no capítulo 2 e com a utilização das bases de dados trabalhadas e descritas no capítulo 3 nomeadas de *posts* e *likes* respectivamente, serão demonstrados os resultados e discussões relativos ao objeto de interesse referente a essa pesquisa, ou seja, descobrir por meio do dos traços de comportamento depressivo e comportamento dos usuários no Facebook, em postagens e/ou curtidas, a probabilidade de um usuário apresentar uma das 4 categorias de sintomas depressivos definidas no BDI-II. Para isso, o modelo de regressão logística multinomial será utilizado. O modelo de regressão logística binária apresentado, servirá com o intuito de realizar um contraste com o modelo multinomial.

4.1 A fase da coleta de dados em discussão

Em pesquisa científica, segundo Marconi e Lakatos [1982] a coleta de dados acontece por meio de técnicas que representam um conjunto de regras ou processos utilizados por uma ciência, neste caso ocorreu por meio do Aplicativo Vivamente desenvolvido especialmente para tal finalidade, conforme mostra a figura 4.1.

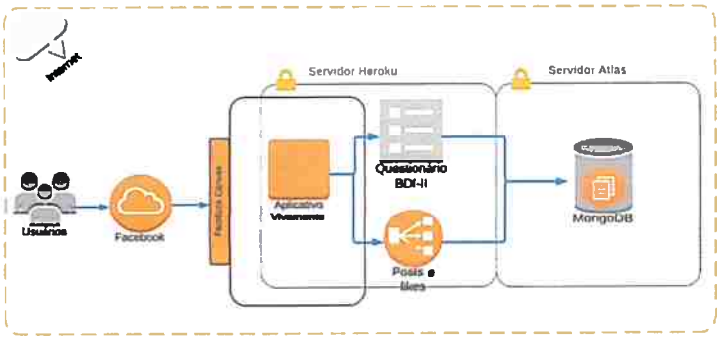


Figura 4.1: *Fluxo da coleta de dados*
Fonte: Elaboração própria

Como observado na figura 4.1, a rede social Facebook foi a fonte para a obtenção dos dados deste estudo. Destaca-se que já é notório e aplicável a utilização das redes sociais para este fim, como destacado nos estudos de Ortigosa *et al.* [2014], Lin e Utz [2015], Shen *et al.* [2015] e Marshall *et al.* [2015] os quais as valorizam e afirmam as mesmas tem um potencial de informações que quando transformadas contribuem para novas descobertas no campo científico em benefício da sociedade.

A coleta dos dados em uma rede social pode ser realizada por diversos métodos e para os mais diversos propósitos. Para pesquisas acadêmicas, segundo Rieder [2013], as três formas de coleta mais utilizadas são: (1) o acesso direto aos servidores de uma empresa, (2) o acesso por meio de *APIs* e, (3) por acesso via *crawler*. Para a etapa de coleta dessa pesquisa o acesso por meio de *APIs* foi o método apropriado. O acesso aos dados via *API* do Facebook foi realizado por meio do desenvolvimento de um *app* específico, essa escolha em detrimento da utilização de dados de repositórios abertos de dados deve-se ao fato do problema

de pesquisa envolver comportamentos na rede social e sintomas depressivos, sendo apenas possível conseguir as duas propriedades por meio de dados não públicos e, dessa forma, houve a necessidade da utilização de um mecanismo seguro e aprovado, que ao mesmo tempo que coletasse os dados na rede, possibilitasse trazer os níveis de sentimentos depressivos de cada um desses usuários.

Portanto, o *app* Vivamente, coletou os dados de postagens e curtidas contidos no histórico de cada usuário da rede social desde o dia que esse usuário começou a fazer parte da rede até o dia da coleta. Além dos dados da rede, também foi obtido o nível de depressão de cada usuário, por meio do BDI-II que foi embutido no mesmo *app*.

Referente aos dados terem sido extraídos das postagens e curtidas, uma pesquisa foi realizada para descobrir quais eram os itens na rede com os quais os usuários mais interagiam e que revelavam comportamentos dos usuários e dentre os trabalhos pode-se citar 2 que afirmam que diversos são os atributos que podem suportar as análises de comportamento do usuário no Facebook, a exemplo das curtidas e postagens que conforme Ryan e Xenos [2011] Kim [2016], são as características mais utilizadas e também onde os usuários passam a maior parte do tempo, respectivamente.

No aspecto relacionado ao Inventário de Depressão de Beck - BDI-II, visando a garantia da precisão das informações referentes aos sintomas depressivos, esse foi considerado um instrumento seguro, por ter sido validado em vários países, inclusive no Brasil, por ser bem conhecido e utilizado em diversas pesquisas na área da psicologia auxiliando em achados bastante relevantes. Isso pode ser demonstrado na afirmação de Sauer *et al.* [2013], ao afirmar que provavelmente o instrumento mais utilizado para medir os sintomas depressivos é o Inventário de Depressão de Beck. Outro exemplo na questão da validação foi o estudo de Nogueira *et al.* [2014], na triagem de depressão em pessoas com epilepsia, onde comprovou a robustez do BDI-II frente aos outros instrumentos, um último exemplo refere-se aos achados de Schutt *et al.* [2016], que para a triagem de sintomas depressivos em candidatos à cirurgia bariátrica, O BDI-II foi uma alternativa viável ao uso do PHQ-9.

Para para Cielen e Meysman [2016], na coleta de dados deve haver a garantia da usabilidade dos dados no projeto, o que significa verificar a existência, qualidade e acesso aos dados, considerando esse conceito, pode-se afirmar que essa etapa foi cumprida com as garantias a ela inerentes.

4.2 A mineração dos dados e níveis de sintomas depressivos

Tanto para Witten *et al.* [2017] quanto para Kumar *et al.* [2009], um processo de mineração de dados segue um determinado conjunto de tarefas para alcançar seus objetivos. Seguindo essa afirmação, na análise dos dados referentes ao segundo objetivo específico, que é preparar os dados coletados pela Escala BDI-II, foram utilizadas técnicas de mineração de dados, especificamente a tarefa de preparação ou pré-processamento dos dados, para categorizar a variável *label4*, conforme expõe a figura 4.2.

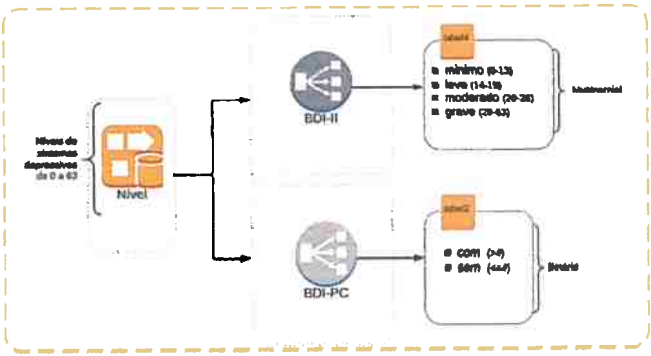


Figura 4.2: *Figura que representa a categorização e binarização da variável nivel*
Fonte: Elaboração própria

Para tornar possível a utilização de um modelo de regressão logística multinomial, bem como seguir a denominação de cortes para o BDI-II, definida por Beck *et al.* [1996] e descrita em Gorenstein *et al.* [2011], a variável que representa o fenômeno de interesse, mantenedora nos 4 níveis de sintomas depressivos dos usuários, denominada por *label4*, precisou ser categorizada.

A escolha na utilização da categorização multinomial ao invés da binária, decorreu do fato de que a variável pôde ser classificada em 4 categorias oriundas de 21 sintomas, o que a tornou mais completa

e robusta, seria admissível reduzir as categorias para duas, mas isso poderia ocasionar alguma perda de informação ou causar imprecisão do que se está tentando averiguar, Allison [2014] afirma essa decisão.

Com relação ao modelo multinomial, essa categorização possibilitou a transformação da variável *label* em 4 variáveis *dummy*, decorrente da quantidade de possibilidades de categorias, isso foi necessário pois as técnicas de regressão logística são utilizadas quando o fenômeno a ser estudado apresenta-se de forma qualitativa, conforme define Fávero e Belfiore [2017].

Essas transformações foram necessárias e fundamentais para que a pesquisa pudesse continuar alinhada aos critérios dos modelos de regressão logística multinomial, transformações semelhantes podem ser destacadas nos trabalhos de Kuramoto *et al.* [2013] que categoriza a variável dependente em 3 categorias de ideação suicida: no ano passado, há mais de um ano, e nunca, Randall *et al.* [2014], categorizou em: sem ideação, com ideação e com ideação e planejamento, Madhu *et al.* [2014], referente a ocorrência de câncer de mama categorizou em alta, média e baixa, Guo *et al.* [2018], referente a causa de trauma na infância, classificou em: sintomas afetivos, psicóticos, ansiosos, maníacos isolados ou com todo o conjunto, Nierop e Germeys [2016] e Kingsbury *et al.* [2018], categorizaram as variáveis pela quantidade de sintomas depressivos presentes: dois, três ou quatro sintomas. A necessidade da pesquisa estar alinhada às especificações da regressão logística, está relacionada com seu objetivo principal que é estimar a probabilidade de ocorrência de um certo nível de comportamento depressivo, considerando o comportamento dos usuários na rede social.

4.3 Extração e polarização das postagens

Referente ao terceiro objetivo, transformar os dados das postagens para o padrão longitudinal e realizar a extração da polaridade dos sentimentos nos textos, essa meta envolve o aumento na qualidade dos dados das postagens e sua preparação para uso em etapas subsequentes, bem como compreende o início da construção do modelo, que conforme Cielen e Meysman [2016] e Schutt *et al.* [2016], está inserida na fase preparação dos dados, esse objetivo está graficamente expresso por meio da figura 4.3.

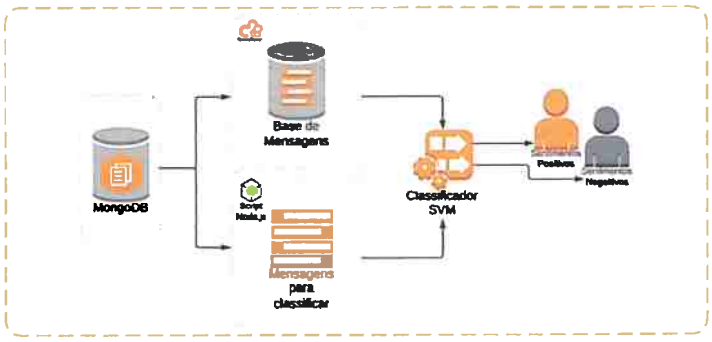


Figura 4.3: *Figura que representa a polarização de sentimentos*
Fonte: Elaboração própria

Os dados textuais obtidos das postagens de cada usuário da rede social Facebook, estão disponíveis, para vários períodos de tempo (várias *cross-sections*, podendo ser observados em horas, dias, semanas, meses, trimestres, anos, etc, portanto, foi possível explorar melhor o potencial desses dados, para tanto, ocorreu a transformação desses dados em formato longitudinal, também chamado de painel, que segundo Fávero e Belfiore [2017], os modelos longitudinais são primordiais nos estudos do comportamento de determinado fenômeno, sendo utilizados em diversas áreas, visto que muitos dados estão disponíveis para várias *cross-sections*.

Para obter maior confiabilidade referente a quantidade de sentimentos positivos e negativos, os dados foram fragmentados até seu último nível, ou seja, até os segundos (ano/mes/dia/hora/minutos/segundos), buscando maior quantidade de informação, maior variabilidade dos dados, menor multicolinearidade, maior número de graus de liberdade e maior eficiência quanto da estimação de seus parâmetros nos sintomas depressivos, que conforme Baltagi [2011] e Wooldridge [2005], são benefícios fornecidos pelos modelos longitudinais de regressão.

Em uma questão mais técnica, porém não menos importante, compete ao mecanismo utilizado para fragmentação dos posts, nesse aspecto, uma query MongoDB foi criada, por questões de compatibilidade com os dados originais armazenados em uma base MongoDB em formato JSON.

Além do propósito da extração dos sentimentos das postagens, o banco de dados transformar-se no padrão longitudinal, será relevante para as análises no próximo objetivo, que considera os períodos de interação do usuário na rede social, com ênfase na estimação da relação entre o nível de depressão e o seu comportamento na rede, os trabalhos de Kuramoto *et al.* [2013] e Kingsbury *et al.* [2018], utilizam-se de dados longitudinais em regressão logística multinomial para estimações semelhantes referentes a ideação suicida.

Após a extração dos textos e suas respectivas datas, foi realizada uma classificação da polaridade dos sentimentos dos textos referentes a de cada uma das postagens, para isso, alguns testes com diferentes algoritmos de classificação de textos, como o algoritmo de *Deep Learning* usando H2O 3.8.2.6, o algoritmo ID3 de árvore de decisão, o *Naive Bayes* e o *SVM* de Stefan Rueping, foram realizados. Algumas variações de configurações foram testadas em cada um dos algoritmos, porém a maior diferença na confiabilidade foi relacionada a etapa de pré processamento de texto.

Os primeiros testes foram realizados com a utilização do dicionário léxico WordNet 3.0, com as etapas de *tokenize, lower case, filter stopwords* e *steeming* e para a extração dos sentimentos foi utilizado o *SentiWordNet 3.0*

A avaliação dos algoritmos de aprendizagem foi realizada por meio da análise de desempenho utilizando Validação Cruzada (*cross validation*), conforme pode ser visto no processo de classificação.

Com a utilização do *dataset WordNet*, obteve-se uma acurácia e precisão de 79,33 e 80%

Como a precisão e a acurácia não estavam satisfatórias, outras alternativas precisaram ser testadas para tentar melhorar esses índices. A primeira decisão foi testar a classificação com uma base de dados de treino específica, até então estava sendo utilizada uma base genérica, em inglês do próprio *WordNet*.

Iniciando pela troca do idioma, duas base de dados em português do Brasil rotuladas com dados oriundos do Twitter foram testadas, uma delas para o domínio político e a outra para classificação de opiniões sobre filmes, tendo sido obtidos acréscimos em média de 4% para a precisão e 5% para a acurácia, indicando que poderia ser um caminho a ser explorado. Sendo assim, a próxima decisão foi a criação de uma base de dados rotulada com sentimentos positivos e negativos com os textos dos próprios atributos *message* de cada objeto do *array posts*.

A base de dados foi rotulada manualmente utilizando a plataforma crowdflower[1] tendo sido gerada uma base de treino rotulada (sentimentos positivos e negativos) com 10.000 sentimentos clasificados em positivos e negativos na proporção de 50% cada polaridade. Essa base passou pela avaliação, primeiramente da psicóloga responsável pelo projeto e posteriormente por um acadêmico do curso de psicologia (colaborador do projeto), auxiliado pela sua orientadora na Universidade Federal de Mato Grosso (UFMT). O *overfitting*, que geralmente é uma preocupação no sentido de que os dados poderiam estar superajustados, não foi levada em consideração nessa fase pois os textos foram classificados para acrescentar mais informações ao *dataset Posts* (nova base de dados desmembrada do *Vivamente*, sendo gerada pelas postagens e desvinculada dos dados das curtidas), que deverá ser utilizada nas próximas etapas deste trabalho.

Em uma nova série de testes, utilizando dessa vez o *dataset Posts*, os resultados foram satisfatórios, chegando a uma precisão de 94,37% , com a utilização do classificador SVM, conforme mostra a figura 3.14.

O classificador H2O utilizando *Deep Learning* obteve uma acurácia e precisão superiores aos do SVM em média 2% , chegando a 96,4, porém a morosidade acabou inviabilizando a continuidade dos testes e sua possível adoção.

Ao final dessa etapa, obteve-se um *datasets* em forma de tabela, chamado *Sentimentos/Message* com 13 atributos desmembrados em datas (ano, mês, dia, hora, minutos, segundos, dia do ano, etc) referentes ao atributo *message* e também com os textos das mensagens classificados em positivos e negativos, representados por um atributo chamado *prediction*, bem como seus níveis de confidencialidade, tanto para os sentimentos positivos quanto para os negativos, representados pelos atributos *confidencialidade(positivo)* e *confidencialidade(negativo)*, respectivamente,

As bases de dados finalizadas, denominadas *posts* e *likes* , estão configuradas em um padrão longitudinal, ou seja, estão dispostas em um formato de painel.

A base de dados *likes* contém 186,911 observações, sendo 76 variáveis independentes e uma dependente. A base de dados *posts* contém 416.663 observações, sendo 52 variáveis independentes e 1 dependente, conforme pode ser visualizado nas tabelas da figura 4.4.

[1] https://www.crowdflower.com

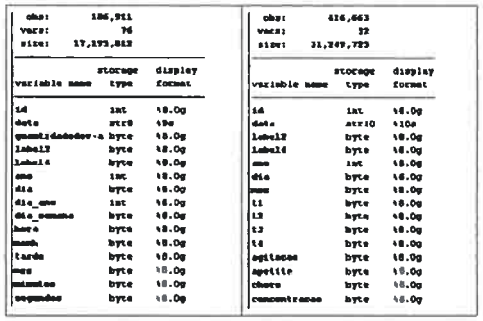


Figura 4.4: *Amostra das características e descrição das variáveis das bases likes e posts*
Fonte: Elaboração própria

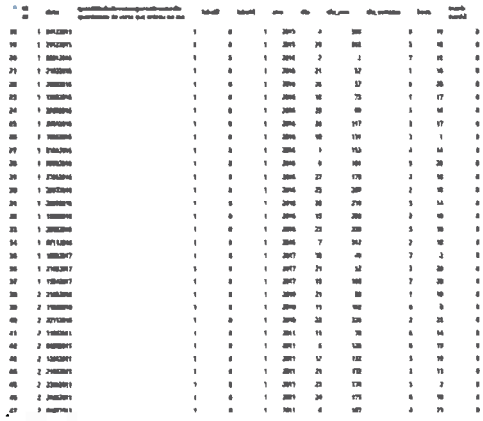


Figura 4.5: *Exemplo de base longa*
Fonte: Elaboração própria

As bases de dados completas podem ser obtidas por meio dos arquivos *base-posts.csv* e *base-likes.csv* no endereço do github[2].

Essa etapa, embora tenha envolvido análise de sentimentos, não pretendeu explorar a sua competência na totalidade, atendo-se apenas a polarização dos sentimentos com o intuito de enriquecer a base de dados para utilização dos testes pertencentes ao modelo de regressão logística multinomial.

4.4 Processo dos experimentos do modelo proposto

Para que o último objetivo pudesse ser alcançado, técnicas de mineração de dados na área da estatísticas foram utilizadas para realização dos experimentos objetivando a predição de probabilidades de traços depressivos, conforme pode ser visualizado na figura 4.6.

[2] http://www.github.com/cmaricy

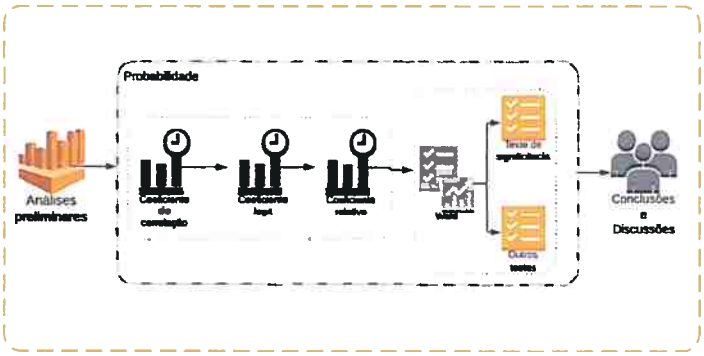


Figura 4.6: *Figura que representa o processo dos experimentos do modelo proposto*
Fonte: Elaboração própria

4.5 Análises preliminares

Algumas medidas de precaução objetivando evitar a violação dos pressupostos de multicolinearidade, heterocedasticidade e autocorrelação foram tomadas e por meio de testes serão analisados e discutidas.

A **Multicolinearidade**, ou seja, indica a existência forte de correlação entre duas (ou mais) variáveis independentes, o teste VIF para verificação de multicolinearidade foi testado para a base de dados likes e posts e seus resultados podem ser verificados na tabela 4.1.

. vif

Variable	VIF	1/VIF
fadiga	2.67	0.373899
energia	2.25	0.444247
desvaloriz~o	2.18	0.459481
prazer	2.04	0.489002
apetite	1.97	0.508420
irritabili~e	1.95	0.513976
concentracao	1.91	0.522796
choro	1.89	0.527725
indecisao	1.83	0.546292
sono	1.73	0.578105
culpa	1.71	0.585613
t2	1.47	0.681198
t4	1.46	0.683003
t1	1.45	0.691321
total_posts	1.41	0.708324
idade	1.33	0.753029
total_likes	1.33	0.753341
madrugada	1.19	0.837252
noite	1.19	0.839780
total_frie~s	1.15	0.873206
sexo	1.10	0.908232
tarde	1.00	0.997601
Mean VIF	1.65	

. vif

Variable	VIF	1/VIF
fadiga	3.27	0.306082
desvaloriz~o	3.05	0.327784
tristeza	2.92	0.342782
energia	2.83	0.353253
interesse	2.82	0.354280
prazer	2.75	0.364030
pessimismo	2.74	0.364360
suicida	2.74	0.364918
fracasso	2.68	0.373049
estima	2.58	0.387311
critica	2.30	0.434475
choro	2.11	0.473764
concentracao	2.09	0.478964
apetite	2.07	0.482122
culpa	2.06	0.484332
irritabili~e	2.06	0.484971
indecisao	1.92	0.521161
sono	1.78	0.560563
idade	1.46	0.685118
total_likes	1.34	0.745238
total_posts	1.30	0.769537
total_frie~s	1.15	0.866030
t1	1.11	0.900456
t2	1.11	0.900630
sexo	1.10	0.907595
Mean VIF	2.13	

Tabela 4.1: *Teste VIF para a verificação de ausência de multicolinearidade*
Fonte: Elaboração própria

Observa-se diante do resultado do teste VIF que todas as variáveis apresentam valores abaixo de 10,

sendo esse valor o padrão exposto na literatura para ausência de multicolinearidade, e estando também em conformidade com a teoria apresentada por Fávero e Belfiore [2017] e descrito na seção 2.3.3, onde afirma que uo valor 4 ainda pode ser considerado alto, porém os valores para os modelos apresentados na figura 4.1, estão abaixo desse valor, podendo-se concluir com segurança a ausência de multicolinearidade entre as variáveis apresentadas.

A **Heterocedasticidade**, que ocorre quando as variâncias não são as mesmas para todas as observações, foi testada por meio do teste de Breusch-Pagan/Cook-Weisberg, que avalia a rejeição ou não da hipótese nula de que os termos de erro sejam homocedásticos, a um determinado nível de significância. A tabela 4.7 mostra o resultado do teste para o modelo, e por meio de sua análise, pode-se perceber que o problema da heterocedasticidade está presente, ou seja, $valor - P\chi^2 = 0,000$, em que o valor de $valor - P\chi^2$ deveria ser maior do que 0,05, necessitando dessa forma que seja executado um método de correção.

Um método de correção bem conhecido para solução desse problema é o chamado estimador robusto Huber/White ou sandwich, de Huber - 1967 e White - 1982, conforme Cora *et al.* [2014] cita em seu estudo, e será utilizado como medida de reparação.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of label4

         chi2(1)      =  4787.95
         Prob > chi2  =   0.0000
```

Figura 4.7: *Teste de Breusch-Pagan/Cook-Weisberg para a verificação de heterocedasticidade*
Fonte: Elaboração própria

Após o modelo ter sido executado com o método de correção robusta de White, novamente foi executado o teste de Breusch-Pagan/Cook-Weisberg para a verificar se o problema de heterocedasticidade foi solucionado. Observa-se diante do resultado do teste exibido na tabela 4.8, que o modelo não mais apresenta o problema de heterocedasticidade apresentado anteriormente.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of label4

         chi2(1)      =  4797.85
         Prob > chi2  =   0.0615
```

Figura 4.8: *Teste de Breusch-Pagan/Cook-Weisberg para a verificação de heterocedasticidade*
Fonte: Elaboração própria

A **Auto-correlação** acontece quando os resíduos da série temporal são autocorrelacionados, e para esse pressuposto o teste *Prais – Winsten* foi executado e seus resultados são apresentado na tabela 4.9.

```
Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =  415612
-------------+------------------------------           F( 25,415586) =       .
       Model |  .000039417    25  1.5767e-06           Prob > F      =  0.0000
    Residual |  2.7843e-06415586  6.6996e-12           R-squared     =  0.9340
-------------+------------------------------           Adj R-squared =  0.9340
       Total |  .000042202415611  1.0154e-10           Root MSE      =  2.6e-06
```

Figura 4.9: *Teste Prais-Winsten para verificação de autocorrelação*
Fonte: Elaboração própria

Observa-se diante do resultado do teste *Prais* que não há problemas de autocorrelação entre as variáveis independentes.

Após os testes que garantem a não violação de alguns pressupostos envolvidos na construção do modelo, outros testes preliminares serão apresentados.

Uma das maneiras pelas quais é possível analisar os dados é por medidas de tendência central (como a média) e de dispersão (como o desvio-padrão). Para investigar possíveis diferenças estatisticamente significativas na quantidade de *posts* em função do estado depressivo, realizou-se um teste *t de student* para

amostras independentes. Nesse teste, assumiu-se que a variável dependente é contínua e normalmente distribuída e a variável independente é categórica e apresenta até dois grupos ou categorias. A versão robusta do teste t, conhecida como *welch test*, lida bem com a violação da homocedasticidade[3] e foi utilizada nessa análise, respaldada pelos trabalhos de Milaniak *et al.* [2018] envolvendo a ocorrência de depressão e estresse em decorrência de resultados negativos em receptores de transplante cardíaco, Jamil [2017], em um sistema automatizado que visa identificar usuários em risco de depressão atividades no Twitter, e Kim [2016] que em sua pesquisa examinou as motivações pertinentes ao check-in no Facebook com relação à preocupação de privacidade, esses 3 trabalhos exemplificados utilizaram o teste t em análises preliminares e com objetivos semelhantesdo teste t aplicado nesse estágio das análises.

Na tabela 4.2 é apresentado, na segunda e quarta colunas, o valor F da estatística e na terceira e quinta colunas o p-valor da estatística bilateral. Para o p-valor inferior a 0,05 é verdade que as médias são diferentes, ou seja, a média da amostra de usuários com sintomas depressivos é diferente da média dos usuários sem sintomas depressivos.

teste t[4]: Comparação[5] de médias para amostra com/sem sintomas depressivos

	Base de dados *posts*		Base de dados *likes*	
Variáveis	**F**	**Sig. (bilateral)**	**F**	**Sig. (bilateral)**
apetite	30.812,77	0	7.119,37	0
choro	28.194,56	0	8.093,85	0
concentração	1.570,56	0	1.824,01	0
culpa	59.233,58	0	20.820,35	0
desvalorização	45.897,11	0	7.213,70	0
domingo	0,376	0,759	-	-
energia	4.546,17	0	1.477,81	0
fadiga	11.344,51	0	8.708,11	0
idade	5.882,57	0	2.725,27	0
indecisão	38.856,58	0	9.491,01	0
irritabilidade	86.435,04	0	29.162,88	0
prazer	36.643,34	0	8.916,09	0
quarta	12,069	0,082	14,189	0,058
sábado	64,608	0	70,128	0
segunda	1,063	0,606	2,278	0,45
sexo	8.112,55	0	6.528,45	0
sexta	3,825	0,328	0,001	0,989
sono	16.851,07	0	8.381,52	0
qtd_message	648,62	0	-	-
qtd_story	1.087,75	0	-	-
qtd_sent_neg	271,25	0	-	-
qtd_sent_pos	23,08	0	-	-
t1	221,382	0	83,439	0
t2	110,474	0	82,924	0
t4	146,252	0	177,221	0
terça	8,093	0,155	-	-
total_dias	5.665,74	0	-	-
total_friends	5.924,03	0	1.186,20	0,008
total_likes	18.093,47	0	8.714,16	0
total_posts	29.493,80	0	3.947,37	0
tarde	-	-	205,491	0
madrugada	-	-	43,587	0,001
noite	-	-	0,179	0,832

Tabela 4.2: *Resultados consolidados referentes ao teste t*
Fonte: Elaboração própria

[3] Conforme Wooldridge [2005] é o termo para designar variância constante dos erros experimentaisε_i para observações distintas.

[4] Resultados gerados no stata 15 e no R Studio

[5] Hipótese testada: as médias das duas amostras, com/sem sintomas depressivos é igual a 0 contra a hipótese de que eles são diferentes de 0, ou seja, H0: b = 0 e H1 b $\neq$ 0

Os resultados consolidados na tabela 4.2, referente ao teste t, apresentam que para a base de dados *posts*, as variáveis *segunda, terça, quarta, sexta* e *domingo* possuem médias iguais, a um nível de confiança de 95%, ou seja, não existe distinção entre a média de interação nesses dias, tanto para usuários com sintomas depressivos quanto para os usuários sem sintomas depressivos.

Por outro lado, as demais variáveis em análise para a base de dados *posts*, apresentam relações significativas entre as médias, indicando que existem diferenças entre os usuários com sintomas depressivos e sem sintomas depressivos.

Para a base de dados *likes*, observa-se na quinta coluna que as variáveis: *noite, sexta, quarta e segunda*, apresentam médias iguais para as amostras de usuários com e sem sintomas depressivos ao nível de 95% de confiança.

Por fim, esses resultados indicam que a interação média dos usuários com e sem sintomas depressivos são iguais nas variáveis temporais, ou seja, os dias de semana e o período de acesso não são distintos em média, o que diferencia essa amostra é o comportamento e os sintomas depressivos do usuário.

4.6 Estimação do modelo de regressão logística multinomial

Antes da estimação dos modelos propriamente ditos, houve a necessidade da definição dos usuários e períodos de tempo, isso foi executado tanto para a base de dados *likes* quanto para a base de dados *posts*, para a variável dependente *label*4.

Primeiramente uma análise geral descritiva foi realizada conforme mostra a saída da figura 4.11.

```
xtset id data1
      panel variable:  id (unbalanced)
       time variable:  data1, 01/01/2010 to 12/16/2017, but with gaps
              delta:  1 day
```

Figura 4.10: *Análise geral descritiva*
Fonte: Elaboração própria

Em que pode-se visualizar que o painel referenciado pela variável *id*, não é balanceado, e a variável que representa o tempo é *data*1, tendo os mesmo resultados para as duas bases de dados.

Na sequência, a distribuição de frequência da variável *label*4 é apresentada na tabela 4.3.

label4	Freq.	Percent	Cum.
1	38,855	20.79	20.79
2	19,411	10.39	31.17
3	23,422	12.53	43.70
4	105,220	56.30	100.00
Total	186,908	100.00	

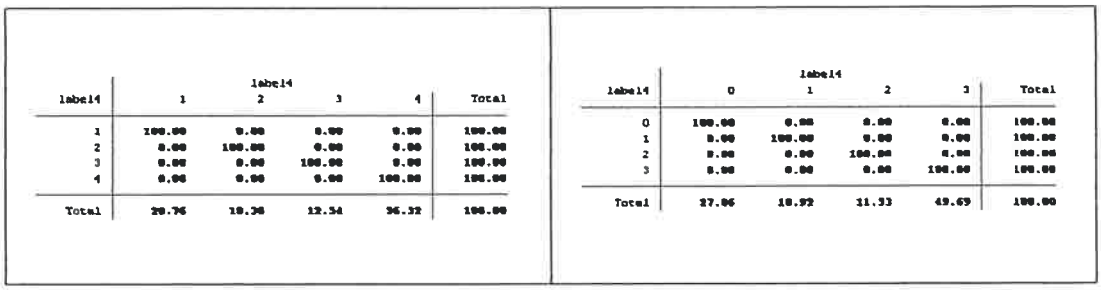
label4	Freq.	Percent	Cum.
0	116,082	27.86	27.86
1	45,481	10.92	38.78
2	48,051	11.53	50.31
3	207,049	49.69	100.00
Total	416,663	100.00	

Tabela 4.3: *Distribuição da frequência da variável label4 para a base likes e posts*
Fonte: Elaboração própria

Onde, pode-se perceber que existem diferenças consideráveis entre o período em que determinado usuário apresentou sintomas depressivos e o período em que não apresentou os sintomas.

Também foi realizada uma investigação de como a variável *label*4 se comporta ao longo do tempo.

As saídas obtidas encontram-se na figura 4.4.



label4	label4				
	1	2	3	4	Total
1	100.00	0.00	0.00	0.00	100.00
2	0.00	100.00	0.00	0.00	100.00
3	0.00	0.00	100.00	0.00	100.00
4	0.00	0.00	0.00	100.00	100.00
Total	20.76	18.38	12.54	36.32	100.00

label4	label4				
	0	1	2	3	Total
0	100.00	0.00	0.00	0.00	100.00
1	0.00	100.00	0.00	0.00	100.00
2	0.00	0.00	100.00	0.00	100.00
3	0.00	0.00	0.00	100.00	100.00
Total	27.06	18.92	11.33	42.69	100.00

Tabela 4.4: *Comportamento de transição da variável label4 para as bases likes e posts*
Fonte: Elaboração própria

Por meio dos resultados apresentados na figura 4.4, é possível verificar que existe considerável persistência do comportamento da variável *label4*, ou seja, analisando o comportamento diário de cada um dos usuários, independente de apresentarem ou não sintomas depressivos, eles tiveram o mesmo comportamento no dia seguinte, isso significa que a presença ou a ausência dos sintomas depressivos não interfere quantitativamente nas ações do usuário.

Elaboradas as análises preliminares, a próxima etapa, refere-se às práticas do modelo de regressão logística, segundo a teoria expressa no Capítulo 2.

4.6.1 Modelo de regressão logística multinomial para a base de dados *likes*

O modelo de regressão que será apresentado leva em consideração a base de dados *likes* e *post* e a variável multinomial *label4*.

O coeficiente de correlação de Pearson é representado pela seguinte equação:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}} \tag{4.1}$$

Onde $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_n$ são os valores medidos de ambas as variáveis, calculado como a razão entre a covariância de duas variáveis e o produto dos desvios-padrão de cada uma delas.

A Tabela 4.5 apresenta os resultados encontrados para o coeficiente de correlação de cada variável com relação a variável *label4* e para uma variável binária (*dummie*) adicionada para identificar cada uma das categorias. Os resultados estão organizados em ordem crescente na categoria grave. O comportamento apresentado pela correlação é relativamente semelhante para as variáveis ao longo das categorias. A variável *idade* apresenta correlação positiva para as categorias mínimo, leve e moderado, com valor igual a 30%, 15% e 7%, respectivamente e coeficiente negativo para a categoria grave, com valor de -39%. De modo que a proporção de quadros mais leves de sintomas depressivos tende a ser maior com o avanço da idade, ao mesmo tempo que os quadros mais graves de sintomas depressivos (categoria grave) são observados entre pessoas mais jovens e tendem a se reduzir com a idade.

Para a variável binária *sexo*, os homens são representados pelo número 1 e as mulheres pelo 0. O coeficiente de correlação estimado para essa variável é relativamente baixo, destacando-se a categoria leve, com coeficiente de 15%, e a categoria grave, com coeficiente de -19%, isto indica que os homens têm probabilidade maior de apresentarem sintomas leves de depressão, ou seja, nos níveis mínimo, leve e moderado, já as mulheres apresentam quadros mais graves de sintomas depressivos. O que corresponde aos relatos de Field [2017], em um estudo longitudinal, na Inglaterra, utilizando regressão logística, onde o *odds ratio* na depressão aos 18 anos foi significativamente maior para as mulheres do que para os homens.

As variáveis[6] t1, t2, t3, t4, segunda, terça, quarta, quinta, sexta, sábado, domingo, tarde, noite, qtd_likes, dia_semana, madrugada, e total_friends apresentam baixos coeficientes de correlação. As variáveis total_likes e *total_posts* possuem valores intermediários para o coeficiente de correlação, -14% e -18% respectivamente, para a categoria mínimo, e 16% e 20% respectivamente, para a categoria grave. Assim, observa-se uma relação negativa entre o número de curtidas e de postagens para os casos mais leves de sintomas depressivos e relação positiva para os casos mais graves, sendo esta uma evidência favorável à argumentação de que o aumento do número de postagens e de curtidas está relacionado à presença de casos mais graves de sintomas depressivos.

[6] As variáveis (t1, t2, t3 e t4) são variáveis temporais, binárias, que correspondem aos trimestres do ano, primeiro, segundo, terceiro e quarto trimestres, respectivamente.

Esta tendência, registrada para as variáveis total_likes e *total_posts* é acentuada para as demais variáveis, cujo coeficiente de correlação assume valores negativos e elevados para a categoria mínimo, coeficientes moderados para as categorias leve e moderado e coeficientes positivos e elevados para a categoria grave.

A estatística descritiva bivariada, também foi utilizada no estudo de Cichowitz *et al.* [2017], onde avaliou a saúde mental em terapia antiretroviral e retenção subseqüente durante seis meses, na África do Sul. Estatísticas descritivas, incluindo proporções para variáveis categóricas e valores médios e faixas interquartis para depressão, transtorno de uso de álcool e retenção no cuidado do HIV. Coeficientes de correlação de Spearman foram calculados entre os instrumentos da pesquisa para avaliar a presença de distúrbios ocorrendo simultaneamente e a co-linearidade entre as variáveis. A exemplo da idade, outras variáveis com p-valor $<$ 0,05 na análise bivariada foram incluídos na regressão logística multivariada, que norteia seu estudo.

Tabela 4.5 - Valor encontrado para o coeficiente de correlação por categoria de sintomas depressivos

	mínimo	leve	moderado	grave	label4
idade	30%	15%	7%	-39%	-39%
sexo	9%	15%	3%	-19%	-17%
manhã	3%	-1%	1%	-2%	-3%
t1	2%	1%	1%	-2%	-2%
t2	1%	1%	0%	-1%	-1%
terça	1%	1%	0%	-1%	-1%
quinta	1%	1%	0%	-1%	-1%
quarta	1%	1%	0%	-1%	-1%
segunda	0%	0%	0%	-1%	-1%
tarde	1%	2%	-2%	-1%	-1%
noite	1%	1%	-1%	-1%	-1%
sexta	0%	0%	0%	0%	0%
qtd_likes	0%	0%	0%	0%	0%
dia_semana	0%	0%	0%	0%	0%
t3	0%	-1%	0%	1%	1%
domingo	-2%	-1%	0%	2%	2%
madrugada	-4%	-1%	2%	2%	4%
sábado	-2%	-1%	0%	2%	2%
t4	-2%	-1%	0%	2%	3%
total_friends	-2%	-5%	-3%	6%	5%
total_likes	-14%	-7%	8%	11%	16%
total_posts	-18%	-5%	-2%	18%	20%
int_sexo	-36%	-9%	-3%	37%	41%
agitação	-39%	-11%	-10%	45%	47%
energia	-57%	-17%	3%	55%	65%
punição	-49%	-18%	-7%	56%	60%
culpa	-46%	-18%	-13%	58%	59%
sono	-48%	-17%	-13%	58%	60%
concentração	-51%	-28%	-3%	60%	65%
irritabilidade	-50%	-20%	-12%	61%	63%
indecisão	-49%	-20%	-15%	62%	63%
tristeza	-54%	-29%	-2%	63%	69%
pessimismo	-52%	-25%	-9%	64%	66%
suicida	-49%	-26%	-12%	64%	65%
apetite	-48%	-24%	-15%	65%	64%
prazer	-58%	-18%	-10%	65%	70%
fadiga	-58%	-24%	-7%	66%	71%
choro	-62%	-14%	-11%	67%	72%
interesse	-58%	-24%	-8%	67%	72%
desvalorização	-62%	-19%	-8%	68%	74%
fracasso	-55%	-26%	-12%	69%	71%
crítica	-48%	-22%	-24%	69%	65%
estima	-59%	-25%	-9%	70%	73%

Tabela 4.5: *Coeficiente de correlação por categoria de sintomas depressivos*
Fonte: Elaboração própria

Fonte: Elaboração própria

A Tabela 4.6 apresenta os coeficientes estimados ln da razão de chance para o modelo *logit multinomial*, com o nível 1 da variável label4 (referência) do modelo logístico, ou seja, as demais categorias são apresentadas em comparação com nível 1 (mínimo) de sintomas depressivos. Os resultados obtidos mostram que todas as variáveis são significativas ao nível de confiança de 95%, as únicas exceções são as variáveis *tarde* e *noite*, que não são significativas para nenhuma das categorias consideradas. A variável *madrugada* não é significativa para a categoria grave, o que indica que as pessoas com quadros mais intensos de sintomas depressivos não apresentam maiores níveis de atividade durante a madrugada, em comparação com as pessoas com menor nível de sintomas depressivos no período da manhã, deixadas na base. Porém, a variável *madrugada* é significativa para as categorias leve e moderado, indicando que as pessoas que apresentam quadros intermediários de sintomas depressivos registram redução de seus níveis de atividade durante a madrugada, em comparação com a base (categoria minimo/referência).

Tabela 4.6 - Valores encontrados para os coeficientes estimados da razão de chance em ln, com o modelo *logit multinomial* para a variável label4.

Variável	leve	p-valor	moderado	pvalor	grave	p-valor
prazer	7,009208	0,000	37,92	0,000	6,646897	0,000
fadiga	3,842227	0,000	28,76	0,000	3,580361	0,000
choro	5,392914	0,000	45,81	0,000	5,162172	0,000
interesse	3,289028	0,000	25,68	0,000	3,038047	0,000
desvalorização	2,98586	0,000	27,03	0,000	2,769329	0,000
fracasso	2,939005	0,000	27,74	0,000	2,731316	0,000
crítica	3,588029	0,000	28,52	0,000	3,341429	0,000
estima	4,34507	0,000	38,14	0,000	4,121762	0,000
concentração	4,052396	0,000	36,6	0,000	3,835375	0,000
irritabilidade	4,754612	0,000	36,42	0,000	4,49875	0,000
indecisão	2,840067	0,000	31,41	0,000	2,662827	0,000
tristeza	0,839303	0,000	5,48	0,762	0,539097	0,000
pessimismo	0,768506	0,000	7,54	0,000	0,568782	0,000
suicida	6,695281	0,000	25,89	0,000	6,188458	0,000
apetite	2,26081	0,000	33,57	0,000	2,128826	0,000
sono	7,334304	0,000	45,81	0,000	7,020534	0,000
culpa	-0,50649	0,000	-3,81	0,000	-0,76697	0,000
energia	1,383766	0,000	15,31	0,000	1,206621	0,000
madrugada	-0,16283	0,050	-1,96	0,006	-0,32582	0,441
tarde	0,067497	0,364	0,91	0,114	-0,07811	0,382
noite	0,012802	0,885	0,14	0,882	-0,16131	0,199
t1	0,19638	0,052	1,94	0,008	-0,00165	0,000
t2	0,053233	0,582	0,55	0,057	-0,13646	0,238
t4	0,232197	0,023	2,28	0,257	0,032215	0,643
total_friends	0,000705	0,000	8,59	0,000	0,000544	0,001
total_likes	-0,00098	0,000	-16,74	0,000	-0,0011	0,000
total_posts	0,001625	0,000	14,27	0,000	0,001402	0,000
sexo	1,326148	0,000	10,09	0,000	1,068608	0,000
idade	-0,04669	0,000	-10,39	0,000	-0,0555	0,110
Cons	-40,7454	0,000	-50,27	0,000	-42,334	0,000
Qui quadrado	411342.57					
Máximoverossimilhança	-7874.3093					
Pseudo R2	0.9631					

Tabela 4.6: *Coeficientes estimados para o modelo logit multinomial*
Fonte: Elaboração própria

Fonte: Elaboração própria

Observa-se que os valores encontrados para os coeficientes estimados da razão de chance em ln, por meio do estimador *logit multinomial*, são bastante informativos, estando de acordo com Abella *et al.* [2017], que utilizou análise multivariada para prever solidão e depressão, e dentre seus achados, expões que o tamanho da rede social está associado com menores chances de solidão, e também encontraram uma associação significativa entre viver em ambiente rural e maiores chances de solidão. Por outro lado, a frequência de

contato, qualidade da rede, nível educacional, status de emprego e renda familiar não estavam associados à solidão. As interações entre marital *status* e depressão, e entre o tamanho da rede social e a depressão foram estatisticamente significantes. Os achados de Kimbrel *et al.* [2016], na história de tentativas de suicídio entre veteranos do Iraque / Afeganistão, a regressão logística demonstrou que a auto lesão não suicida*Non Suicidal Self Injury* (NSSI) permaneceu um preditor significativo de tentativas de suicídio, mesmo após uma ampla gama de covariáveis.

Após os resultados estimados pelo modelo *logit multinomial* é possível estimar a probabilidade condicional de ocorrência dos diferentes níveis de sintomas depressivos em relação às variáveis em análise. Neste caso, a função de probabilidade é expressa pela seguinte equação:

$$p_i = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-(\alpha+\beta_1.X_{1i}+\beta_2.X_{2i}+...+\beta_k.X_{ki})}} \tag{4.2}$$

A probabilidade condicional de ocorrência de sintomas depressivos dos níveis em análise (leve, moderado e grave) em comparação com o nível mínimo (base/referência) é dada pela expressão que varia entre o intervalo (0 e 1). Desta forma, é possível calcular a probabilidade esperada do evento. Considerando um usuário com 25 anos, do sexo masculino, com 35 postagens e 98 curtidas no total, com 191 amigos, que está interagindo no quarto trimestre e no turno[7]da madrugada, a escala de sono e apetite é 1, e as demais variáveis em análise tendo peso 0, ou seja, não constam[8] para o usuário, assim, a probabilidade deste usuário ter comportamento depressivo leve é igual a 0, pois, substituindo os valores estimados para a categoria leve e os dados observados do usuário na função de probabilidade condicional dado pelo modelo logístico na equaçã 4.2 apresenta:

$$p_i = \frac{1}{1+e^{-(-30,826)}} \cong 0 \tag{4.3}$$

Considerando os valores estimados na Tabela 4.6 e aplicando a exponencial, são apresentados na Tabela 4.7 as razões de chance. Observa-se que a chance de um usuário que acessa a rede social de madrugada se encontrar na categoria leve é de 85% comparado a categoria mínimo. Ou seja, os usuários da categoria leve possuem probabilidade 15% menor do que os usuários da categoria mínimo de acessarem a rede social de madrugada. De modo semelhante, os usuários da categoria moderado possuem 24,9% de probabilidade a menos do que os usuários da categoria mínimo de acessar a rede social de madrugada.

Em relação aos resultados encontrados para os períodos do ano, os coeficientes estimados para o primeiro trimestre, apresentados na Tabela 4.7, são significativos para todas as categorias, indicando que, independente do quadro de comportamento depressivo, todas as pessoas tendem a acessar mais sua rede social no primeiro trimestre, mostrando que, em média, as pessoas da categoria leve apresentam 21,7% (121,7% -100%) de probabilidade a mais de acessarem a rede social no primeiro trimestre do que as pessoas da categoria mínimo. As pessoas da categoria moderado apresentam 47,1% a mais de probabilidade de utilizar a rede social no primeiro trimestre em comparação com as pessoas da categoria mínimo. Já as pessoas da categoria grave apresentam 198,7% mais de probabilidade de utilizar a rede social do que as pessoas da categoria mínimo, no primeiro trimestre.

Os coeficientes estimados para o segundo trimestre do ano são significativos apenas para a categoria moderado, indicando que apenas as pessoas desta categoria apresentam maior chance do que as pessoas da categoria mínimo de utilizar a rede social neste trimestre. O valor encontrado para o coeficiente mostrado na Tabela 4.6, é de 0,55, com p-valor de 0,057. De fato, a Tabela 4.7 mostra que as pessoas da categoria leve apresentam 24% de probabilidade menor de acessarem a rede social no segundo trimestre do que as pessoas da categoria mínimo.

Os coeficientes estimados para o quarto trimestre mostram na tabela 4.6 que apenas a categoria leve apresenta coeficiente significativo, com coeficiente estimado de 0,232 e p-valor igual à 0,023. A Tabela 4.7 mostra que as pessoas da categoria leve apresentam 26,1% de probabilidade a mais de acessarem a rede social no quarto trimestre do que os usuários da categoria mínimo.

Assim, os resultados encontrados para os trimestres mostram que todas as pessoas utilizam mais a rede social no primeiro trimestre, mas que a probabilidade de uso se eleva com o nível de sintomas depressivo que a pessoa apresenta. Por outro lado, no segundo e no quarto trimestres não se observa uma relação direta entre o uso da rede social e a gravidade do quadro de sintomas depressivos.

O teste de Wald é utilizado para identificar se os usuários da categoria moderado e da categoria grave utilizam a rede social com a mesma chance no primeiro trimestre do ano. O valor encontrado para este teste foi de 1,742, com p-valor de 0,000, indicando que os usuários da categoria grave utilizam mais a rede social do que os usuários da categoria moderado, no primeiro trimestre. Este mesmo teste foi utilizado para

[7] Turno de horas: (manhã >6; <=12) ;(tarde >12; <=18); (noite >18;<=24); (madrugada <24;<=6 h)

[8] Os dados não foram observados, ou seja, o usuário não possui.

identificar se os usuários da categoria moderado utilizam mais a rede social no primeiro trimestre do que os usuários da categoria leve. O resultado encontrado foi de 3,29, com p-valor de 0,070, isto mostra que os usuários da categoria moderado realmente utilizam mais a rede social. Porém, a diferença não é tão elevada quanto a observada entre a categoria grave e a categoria moderado, pois o valor estimado para o teste só é significativo quando se considera um nível de confiança[9] de 90%.

Os resultados encontrados para os coeficientes estimados para o modelo *logit multinomial* exibidos na Tabela 4.6, para a interação na rede social indicam que as pessoas que se encontram na categoria leve apresentam maior número de amigos do que as pessoas da categoria mínimo, pois o sinal do coeficiente estimado é positivo e significativo. A Tabela 4.7 mostra que a probabilidade do usuário da categoria leve em acessar a rede social em relação à categoria mínimo, se eleva em 0,1% com o acréscimo de um novo amigo (100,1%-100%).

A Tabela 4.7 também evidencia que a probabilidade de uma pessoa da categoria leve acessar a rede social se reduz em 0,2% (99,8% - 100%), em relação a uma pessoa da categoria mínimo, se ela possuir mais um amigo. E, a probabilidade de uma pessoa da categoria grave acessar as redes sociais, em comparação com a categoria mínimo, se reduz em 0,1% (99,9% - 100%) com a presença de mais um amigo. De modo que o acréscimo no número de amigos desmotiva as pessoas com maior nível de comportamento depressivo a acessar a rede social.

O sinal do coeficiente estimado para o total de curtidas indica que as pessoas das categorias leve, moderado e grave curtem menos páginas. A Tabela 4.7 mostra que a probabilidade de uma pessoa da categoria leve acessar a rede social se reduz em 0,1% (99,9%-100%) caso ela curta um uma nova publicação, em comparação com os usuários da categoria mínimo. A probabilidade das pessoas da categoria moderado em acessar a rede social se reduz em 0,4% (99,6%-100%) caso ela conceda uma nova curtida. E, a probabilidade de uma pessoa da categoria grave acessar a rede social se reduz em 0,3% (99,7% - 100%). De modo que as pessoas com casos mais leves de comportamento depressivo são mais propensas a acessar com maior frequência a rede social caso ele curta novas publicações do que as pessoas com casos mais graves de comportamento depressivo, com destaque para a categoria moderado.

Todavia, a probabilidade dos usuários da categoria leve acessarem a rede social é 0,2% (100,2% - 100%) superior ao ser observado para as pessoas da categoria mínimo, caso eles efetivem uma nova postagem. De modo semelhante, as probabilidades dos usuários das categorias moderado e grave em acessar a rede social se eleva em 0,6%, em relação à categoria mínimo, caso eles realizem uma nova postagem. Assim, os usuários com maiores níveis de comportamento depressivo são mais propensos a acessar a rede social após realizarem novas postagens.

De maneira semelhante, Kingsbury *et al.* [2018], fundamentado em bons parâmetros de ajuste na trajetória de sintomas depressivos e resultados estatisticamente significativos, utilizou um modelo de regressão logística multinomial que consistiu de três trajetórias de sintomas depressivos, baixa, moderada-estável, moderada-crescente, ao longo de um período de 27 anos. Tendo alcançado o objetivo do estudo, que foi investigar se alguns fatores adversos previam trajetórias de sintomas depressivos ao longo de um período de 27 anos após o nascimento de um bebê.

Mufudza e Erol [2016], também utilizou modelos de regressão, afirmando que a predição da doença cardíaca pode eficazmente identificar os principais riscos usando um modelo de regressão. O modelo ajuda na inferência nas diferentes categorias não somente dentro da amostra, mas também dentro de cada uma das categorias.

Tabela 4.7 - Valor encontrado para os coeficientes transformados para relativos (razão de chances) para a variável dependente label4, que representa os 4 níveis de comportamento depressivo.

[9] Na literatura utiliza-se 95% e 90% (se o p-valor é menor que 0,01 afirmamos que ele é significativo a 99%, caso seja, 0,05 afirmamos que ele é significativo a 95% e a 0,1 significativo a 90%). Usualmente, quando o coeficiente é significativo a 99% falamos que o valor é significativo a um nível de superior ou igual a 95%.

Variável	Leve	Moderado	Grave
prazer	1106,778	414927,880	82239805,941
fadiga	46,629	7082,569	25372,216
choro	219,843	18705,820	457430,390
interesse	26,817	50121,233	33220908,176
desvalorização	19,804	42497,477	45261167,806
fracasso	18,897	3622,215	1952653,940
crítica	36,163	21,505	940,925
estima	77,097	3482,665	1185980,006
concentração	57,535	94093,100	225650973,854
irritabilidade	116,119	3349896,474	1721737857,592
indecisão	17,117	1364,643	2011859,547
tristeza	2,315	1,068	5365821,262
pessimismo	2,157	10785,339	222566,230
suicida	808,581	686402,864	2059121,513
apetite	9,591	25,235	33772,090
sono	1531,961	791353,461	109618860,223
culpa	0,603	379249,660	2725778738,958
energia	3,990	106,780	1338,421
madrugada	0,850	0,716	0,889
tarde	1,070	0,837	1,130
noite	1,013	1,020	1,233
t1	1,217	1,471	2,987
t2	1,055	0,760	0,810
t4	1,261	1,180	1,087
total_friends	1,001	0,998	0,999
total_likes	0,999	0,996	0,997
total_posts	1,002	1,006	1,006
sexo	3,767	0,026	10,403
idade	0,954	1,235	1,033
Cons	0,000	0,000	0,000

Tabela 4.7: *Valor encontrado para os coeficientes transformados para relativos*
Fonte: Elaboração própria

A Tabela 4.8 apresenta os resultados obtidos para o teste de significância conjunta de Wald, para cada variável estimada. Este teste identifica se cada uma das variáveis que se encontra no modelo explica, em termos estatísticos, os valores observados para a variável label4 (quatro níveis de sintomas depressivos), sem a discriminação dos dados em diferentes categorias. Além do teste de significância conjunta de cada uma das variáveis, a Tabela 4.8 também apresenta dois testes adicionais. O primeiro identifica se as variáveis total_friends, total_likes e total_posts explicam, conjuntamente, a variável label4. O segundo teste verifica se os trimestres do ano são significativos conjuntamente[10].

Os resultados obtidos mostram que os valores calculados para todos os testes são significativos ao nível de confiança de 99% (pvalor igual a 0,00). Deste modo, há evidências de que cada uma das variáveis adicionadas ao modelo explica o comportamento da variável label4, de um modo geral, sem que ela seja discriminada em categorias. Ademais, as variáveis relacionadas ao agrupamento (total_friends, total_likes e *total_posts*) e ao trimestre do ano também são significativas, indicando que existe uma relação estatística entre a interação do usuário com a rede social e o período do ano com as diferentes categorias de comportamento depressivo identificadas pela variável label4. O mesmo é válido para as variáveis *idade* e *sexo*, existindo evidências de que a probabilidade de acesso à rede social é diferente entre as faixas etárias e o sexo.

Nierop e Germeys [2016], utilizou Regressão logística multinomial para avaliar se o trauma na infância estava mais fortemente associado com alguns sintomas afetivos isolados ou em conjunto. Tendo alcançado seu objetivo, descobrindo que o trauma estava consideravelmente mais fortemente associado com o conjunto de sintomas, e não com esses sintomas de forma individual.

Ophir *et al.* [2017], utilizou uma análise de regressão múltipla que revelou quatro atributos na atualização de status no Facebook na predição os escores de depressão: (1) sintomas depressivos do DSM-5 (incluindo emocional e comportamental, não sintomas somáticos); (2) distorções cognitivas; (3) forma poético-dramática no conteúdo verbal; e (4) atitudes para com os outros. Aplicou uma análise de regressão múltipla, para testar

[10] Os dias da semana não foram adicionados ao modelo pois os coeficientes estimados não foram significativos.

quais características contribuem mais para a previsão dos escores de avaliação da depressão.

Tabela 4.8 - Teste de significância conjunta de Wald.

Variável	Wald	p-valor	Variável	Wald	pvalor
prazer	2528.20	0,00	**sono**	3138.25	0,00
fadiga	1742.05	0,00	**culpa**	1614.20	0,00
choro	3386.52	0,00	**energia**	754.45	0,00
interesse	1635.31	0,00	**madrugada**	12.99	0,00
desvalorização	2389.10	0,00	**t1**	48.32	0,00
fracasso	2240.15	0,00	**t2**	10.04	0,00
crítica	1399.39	0,00	**t4**	6.18	0,00
estima	2450.43	0,00	**total_friends**	457.55	0,00
concentração	2684.15	0,00	**total_likes**	1118.79	0,00
irritabilidade	2710.08	0,00	**total_posts**	600.03	0,00
indecisão	2416.38	0,00	**sexo**	889.12	0,00
tristeza	943.95	0,00	**idade**	694.51	0,00
pessimismo	825.50	0,00	**total_friends**	1754.73	0,00
suicida	1387.79	0,00	**total_likes**	1754.73	0,00
apetite	2254.30	0,00	**total_posts**	1754.73	0,00
t1 t2 t3	90,97	0,00			

Tabela 4.8: *Teste de significância conjunta de Wald*
Fonte: Elaboração própria

Além do teste de Wald de significância conjunta das variáveis, apresentado na Tabela 4.8, também foram realizados alguns testes de Wald[11]adicionais, com o objetivo de identificar se as variáveis relacionadas à interação por meio do Facebook se modificam entre as categorias de sintomas depressivo, conforme mostra a Tabela 4.9, onde a segunda e a quarta colunas apresentam a hipótese nula testada e consolidam o valor encontrado para o teste de Wald e a terceira e a quinta colunas vislumbram os resultados encontrados para o p-valor.

Os resultados obtidos para o teste de Wald, com o objetivo de identificar se os coeficientes estimados para a categoria leve são diferentes dos estimados para a categoria moderado, mostram que apenas as variáveis *noite* e *t4* apresentam coeficientes iguais em termos estatísticos (p-valor superior a 0,10). E, as variáveis *madrugada* e *t1* apresentam coeficientes diferentes apenas quando o nível de confiança é relaxado de 1% para 10%. Todas as demais variáveis são significativas ao nível de confiança de 99%, indicando que os coeficientes estimados para a categoria leve são diferentes dos estimados para a categoria moderado. Isto é, a influência destas variáveis sobre o acesso a rede social se modifica em resposta aos níveis de sintomas depressivos.

Já o teste de Wald que verifica se os coeficientes estimados para a categoria grave são iguais, em termos estatísticos, aos estimados para a categoria moderado indica que as variáveis t4 e total_posts, apresentam coeficientes iguais ao nível de confiança de 99%. Já a variável noite apresenta coeficientes iguais apenas quando o nível de confiança é relaxado de 1% para 5% (p-valor igual a 0,442).

Tabela 4.9 - Teste de Wald para as categorias

[11] O teste de Wald testa as hipóteses, que os coeficientes são iguais a 0 contra a hipótese de que eles são diferentes de 0, ou seja, H0: b = 0 e H1 b $\neq$ 0

Variável	moderado = leve	pvalor	grave = moderado	pvalor
prazer	597.84	0.000	496.41	0.000
fadiga	812.30	0.000	102.19	0.000
choro	551.81	0.000	738.67	0.000
interesse	596.77	0.000	378.00	0.000
desvalorização	1010.44	0.000	650.67	0.000
fracasso	745.37	0.000	729.20	0.000
crítica	20.58	0.000	565.26	0.000
estima	553.35	0.000	437.79	0.000
concentração	825.76	0.000	500.23	0.000
irritabilidade	561.26	0.000	827.50	0.000
indecisão	555.85	0.000	874.81	0.000
tristeza	25.00	0.000	888.54	0.000
pessimismo	427.72	0.000	337.47	0.000
suicida	623.23	0.000	85.42	0.000
apetite	160.42	0.000	961.28	0.000
sono	322.53	0.000	717.14	0.000
culpa	1100.29	0.000	499.93	0.000
energia	293.42	0.000	223.63	0.000
madrugada	3.65	0.056	5.50	0.019
tarde	8.37	0.003	12.91	0.000
noite	0.00	0.945	3.90	0.048
t1	3.29	0.069	41.25	0.000
t2	9.37	0.002	0.37	0.545
t4	0.41	0.523	0.59	0.442
total_friends	173.30	0.000	210.17	0.000
total_likes	676.46	0.000	145.66	0.000
total_posts	395.59	0.000	1.67	0.196
sexo	314.26	0.000	470.80	0.000
idade	480.39	0.000	122.53	0.000

Tabela 4.9: *Teste de Wald para as categorias*
Fonte: Elaboração própria

4.6.2 Modelo de regressão logística multinomial considerando a base de dados *posts*

Os resultados encontrados para a base de dados *posts* encontram-se consolidados na tabela 4.10. A comparação com a base de dados *likes* mostra que os valores estimados para o primeiro e o segundo trimestres, para a categoria moderado, não são significativos para a base de dados *posts*, ao passo que eram significativos para a base de dados *likes*, de modo semelhante, para o quarto trimestre, a categoria grave passa a apresentar coeficientes significativos. Assim, a categoria moderado deixa de ser sensível à influência exercida pelo primeiro e segundo trimestres e a categoria grave passa a ser influenciada pelo primeiro e pelo quarto trimestres.

Em relação às variáveis que representam os agrupamentos, a única modificação observada, em termos de significância estatística, é para a variável *total_posts*, que deixa de ser significativa para a categoria grave. De modo que a realização de novas postagens não influencia na probabilidade de acesso para os usuários com níveis mais elevados de sintomas depressivos. Ademais, o coeficiente estimado para a variável *tristeza*, para a categoria moderado, também deixa de ser significativo.

Tabela 4.10 Valor encontrado para os coeficientes estimados da razão de chance em *ln*, com o modelo *logit multinomial* para a variável *label*4.

Variável	leve	pvalor	moderado	pvalor	grave	pvalor
prazer	1,72E+06	0,000	6,99E+08	0,000	1,46E+12	0,000
fadiga	988,557	0,000	6,14E+05	0,000	1,35E+06	0,000
choro	1807,092	0,000	1,95E+05	0,000	1,05E+07	0,000
interesse	2,635	0,000	7025,765	0,000	2,22E+08	0,000
desvalorização	30,566	0,000	60269,850	0,000	2,55E+09	0,000
fracasso	11,540	0,000	4999,933	0,000	2,86E+07	0,000
crítica	2367,393	0,000	9316,062	0,000	3,72E+06	0,000
estima	125,440	0,000	2538,730	0,000	8,72E+07	0,000
concentração	2099,187	0,000	2,92E+05	0,000	6,15E+10	0,000
irritabilidade	74,264	0,000	2,50E+05	0,000	3,55E+09	0,000
indecisão	21,007	0,000	443,152	0,000	6,13E+07	0,000
tristeza	95,965	0,000	2,461	0,000	5,31E+10	0,000
pessimismo	42,660	0,000	6,85E+06	0,000	4,33E+08	0,000
suicida	94,956	0,000	3107,153	0,000	3,23E+04	0,000
apetite	32,810	0,000	1045,259	0,000	2,26E+08	0,000
sono	69,874	0,000	11439,620	0,000	7,16E+07	0,000
culpa	103,838	0,000	2,21E+06	0,000	1,28E+13	0,000
energia	2,607	0,000	433,348	0,000	4,25E+04	0,000
t1	0,914	0,082	1,013	0,887	0,782	0,040
t2	0,940	0,219	0,985	0,864	0,917	0,469
t4	0,915	0,075	0,877	0,129	0,730	0,007
total_friends	1,002	0,000	1,000	0,000	1,002	0,000
total_likes	1,002	0,000	1,006	0,000	1,006	0,000
total_posts	0,997	0,000	0,997	0,000	1,000	0,715
sexo	0,155	0,000	0,000	0,000	12,532	0,000
idade	0,963	0,000	1,146	0,000	0,841	0,000
Cons	0,000	0,000	0,000	0,000	0,000	0,000
Qui quadrado		9,53E+05				
Máximoverossimilhança		- 2,02E+04				
PseudoR^2		0,959				

Tabela 4.10: *Coeficientes estimados para o modelo logit multinomial*
Fonte: Elaboração própria

Quanto ao valor dos coeficientes para a base de dados *posts*, a Tabela 4.11 mostra que, para o primeiro e o quarto trimestres a probabilidade de acesso à rede social para as categorias leve e grave eram superiores ao observado para a categoria mínimo, em comparação com a base de dados *likes*, e se tornam inferiores. A probabilidade de um usuário da categoria leve acessar a rede social no primeiro trimestre é 8,6% (91,4% - 100%) inferior ao observado para a categoria mínimo. Resultado semelhante é observado para a categoria grave, cuja probabilidade de acesso a rede social no primeiro trimestre se torna 21,8% (78,2% - 100%) inferior ao observado para a categoria mínimo. Para o quarto trimestre, a probabilidade de um usuário da categoria leve acessar a rede social é 8,5% (100% - 91,5%) inferior ao observado para a categoria mínimo, já a probabilidade de um usuário da categoria grave acessar a rede social é 27,1% (100% - 72,9%) inferior ao registrado para a categoria mínimo.

As variável que representa o total de amigos mostra que o número de amigos exerce efeito mais alto sobre a probabilidade de acesso do que o observado para a base de dados *likes*. Os usuários da base de dados *posts* acessam mais a rede social quando se tornam amigos de um novo usuário. A probabilidade de um usuário da categoria leve acessar a rede social ao acrescentar um novo amigo é 0,2% (100,2% - 100%) superior a um usuário da categoria mínimo, a probabilidade de um usuário da categoria moderado é 0,4% (100,4% - 100%) superior ao observado para um usuário da categoria mínimo e a probabilidade de um usuário da categoria grave é 0,2% (100,2% - 100%) superior.

A comparação do efeito do total de curtidas sobre a probabilidade de acesso para as bases de dados *likes* e *posts* mostra que os usuários da base de dados *likes* possuem maior propensão a acessar a rede social após efetuarem uma nova curtida, do que os usuários da base de dados *posts*. A probabilidade de um usuário da categoria leve acessar a rede social quando efetua um novo *like* é 0,2% (100,2% - 100%) superior a um usuário da categoria mínimo. Já as probabilidades de um usuário da categoria moderado e da categoria grave são 0,6% (100,6% - 100%) superiores.

Para a variável relativa ao total de postagens, a comparação entre os coeficientes estimados para as bases de dados *likes* e *posts*, revela que a probabilidade dos usuários desta última base de dados acessarem a rede social se torna inferior ao observado para usuários da categoria mínimo, após realizarem uma nova postagem. Um usuário da categoria leve possui 0,3% (99,7% - 100%) de probabilidade a menos do que um usuário da categoria mínimo. Já um usuário da categoria moderado apresenta probabilidade 0,3% inferior.

Em suma, os usuários da base de dados *posts* possuem maior probabilidade de acessarem a rede social após obterem novos amigos e novas curtidas, mas apresentam menor probabilidade de acesso após realizarem novas postagens. A probabilidade de acesso de usuários com níveis mais elevados de sintomas depressivos é superior ao registrado para a categoria mínimo após obterem novos amigos e novas curtidas e inferior após realizarem novas postagens.

Tabela 4.11 - Valor encontrado para os coeficientes transformados para relativos (*razão de chances*) para a variável dependente *label*4.

Variável	Leve	Moderado	grave
prazer	1716843	6,99E+08	1,46E+12
fadiga	988,5565	614359,5	1350290
choro	1807,092	194975,4	1,05E+07
interesse	2,634501	7025,765	2,22E+08
desvalorização	30,56627	60269,85	2,55E+09
fracasso	11,53974	4999,933	2,86E+07
crítica	2367,393	9316,062	3718347
estima	125,4396	2538,73	8,72E+07
concentração	2099,187	292488	6,15E+10
irritabilidade	74,26389	249766,1	3,55E+09
indecisão	21,00693	443,1518	6,13E+07
tristeza	95,96531	2,460643	5,31E+10
pessimismo	42,66038	6849052	4,33E+08
suicida	94,95551	3107,153	32334,65
apetite	32,80997	1045,259	2,26E+08
sono	69,87429	11439,62	7,16E+07
culpa	103,8383	2206244	1,28E+13
energia	2,607392	433,3478	42464,24
t1	0,914264	1,012699	0,781934
t2	0,94006	0,985216	0,917284
t4	0,915204	0,877187	0,729508
total_friends	1,002204	1,000454	1,001633
total_likes	1,002306	1,006206	1,005929
total_posts	0,997164	0,997116	0,999955
sexo	0,155106	9,81E-05	12,53176
idade	0,963363	1,145511	0,841211
Cons	4,28E-24	9,34E-66	7,90E-164

Tabela 4.11: *Valor encontrado para os coeficientes transformados para relativos*
Fonte: Elaboração própria

A fim de comparar os resultados entre as categorias, o teste de Wald é aplicado para identificar se a probabilidade de acesso a categoria moderado é superior à probabilidade de acesso da categoria leve e se a probabilidade de acesso da categoria grave é superior à probabilidade de acesso da categoria moderado. Conforme observado na Tabela 4.12, o teste de Wald indica que os coeficientes estimados para a categoria moderado são diferentes dos estimados para a categoria leve, as únicas exceções são os resultados estimados para os trimestres e para o total de postagens, para os quais não se pode rejeitar a hipótese nula de igualdade nos coeficientes estimados. O p-valor estimado é superior a 0,05, indicando que os valores obtidos para o teste não são significativos ao nível de confiança de 95%.

O valor encontrado para o teste de Wald, para a comparação entre os coeficientes estimados para as categorias moderado e grave retornou valores significativos para todos os testes aplicados. A única exceção é o segundo trimestre. De modo que, salvo para esta variável, rejeita-se a hipótese nula de igualdade entre os coeficientes estimados para estas duas categorias.

Tabela 4.12 - Valores encontrados para o teste de Wald discriminado por categoria.

Variável	Wald	p-valor	Wald	p-valor
prazer	3482,22	0,0000	1603,31	0,0000
fadiga	2572,58	0,0000	73,29	0,0000
choro	1917,74	0,0000	1239,46	0,0000
interesse	1975,61	0,0000	1338,80	0,0000
desvalorização	3872,61	0,0000	2546,54	0,0000
fracasso	3480,36	0,0000	2324,34	0,0000
crítica	272,38	0,0000	2445,56	0,0000
estima	1353,81	0,0000	1949,01	0,0000
concentração	2085,04	0,0000	1789,02	0,0000
irritabilidade	2105,92	0,0000	1794,20	0,0000
indecisão	1544,30	0,0000	2122,59	0,0000
tristeza	1343,26	0,0000	2525,22	0,0000
pessimismo	1638,25	0,0000	1156,14	0,0000
suicida	789,78	0,0000	736,24	0,0000
apetite	1237,03	0,0000	2599,16	0,0000
sono	1575,21	0,0000	1852,89	0,0000
culpa	3539,49	0,0000	1794,91	0,0000
energia	1457,57	0,0000	976,22	0,0000
t1	2,00	0,1571	10,28	0,0013
t2	0,44	0,5063	0,76	0,3833
t4	0,36	0,5471	5,49	0,0191
total_friends	415,12	0,0000	503,88	0,0000
total_likes	1889,59	0,0000	5,34	0,0208
total_posts	0,77	0,3809	1018,11	0,0000
sexo	2504,70	0,0000	1606,64	0,0000
idade	758,12	0,0000	474,72	0,0000

Tabela 4.12: *Valores encontrados para o teste de Wald discriminado por categoria*
Fonte: Elaboração própria

Por fim, uma segunda versão do teste de Wald é estimada para identificar se todas as variáveis são significativas, conjuntamente, para todas as categorias consideradas. Os resultados encontrados na Tabela 4.13, mostram que todas as variáveis são significativas ao nível de confiança de 95%, sendo o segundo trimestre a única exceção, gerando assim, evidências de que todas as variáveis, exceto o segundo trimestre, influenciam no acesso a rede social.

A Tabela 4.13, também mostra os resultados do teste que se as variáveis de agrupamento, (total_friends, total_likes, total_posts) e trimestrais são conjuntamente significativas. Ambos os testes indicam que estes dois conjuntos de variáveis são significativos. De modo que rejeita-se a hipótese nula de que elas, em grupo, não influenciam no acesso a rede social.

Tabela 4.13 - Teste de Wald para a significância conjunta das categorias.

Variável	wald	pvalor	Variável	wald	pvalor
prazer	7910,88	0,0000	apetite	6452,25	0,0000
fadiga	4917,17	0,0000	sono	5466,39	0,0000
choro	5927,81	0,0000	culpa	7215,98	0,0000
interesse	3546,06	0,0000	energia	2604,10	0,0000
desvalorização	8503,01	0,0000	t1	15,30	0,0000
fracasso	6829,57	0,0000	t2	2,71	0,4386
crítica	5245,46	0,0000	t4	9,03	0,0289
estima	6180,92	0,0000	total_friends	1920,73	0,0000
concentração	6811,51	0,0000	total_likes	2832,52	0,0000
irritabilidade	5706,46	0,0000	total_posts	2817,73	0,0000
indecisao	5903,93	0,0000	sexo	4444,38	0,0000
tristeza	5011,04	0,0000	idade	1390,90	0,0000
pessimismo	4232,34	0,0000	total_friends, total_likes, total_posts	6019,53	0,0000
suicida	2298,10	0,0000	t1, t2, t3	20,98	0,0128

Tabela 4.13: *Teste de Wald para a significância conjunta das categorias*
Fonte: Elaboração própria

4.7 Estimação do modelo de regressão logística binária

Antes da estimação dos modelos propriamente ditos, houve a necessidade da definição dos usuários e períodos de tempo, isso foi executado tanto para a base de dados *likes* quanto para a base de dados *posts*, para a variável dependente *label4*.
Primeiramente uma análise geral descritiva foi realizada conforme mostra a saída da tabela 4.11.

```
xtset id data1
      panel variable:  id (unbalanced)
       time variable:  data1, 01/01/2016 to 12/16/2017, but with gaps
               delta:  1 day
```

Figura 4.11: *Análise geral*
Fonte: Elaboração própria

Em que pode-se visualizar que o painel referenciado pela variável *id*, não é balanceado, e a variável que representa o tempo é *data*.
Na sequência, a distribuição de frequência da variável *label4* é apresentada na tabela 4.14.

label2	Freq.	Percent	Cum.
0	50,697	27.12	27.12
1	136,211	72.88	100.00
Total	186,908	100.00	

label2	Freq.	Percent	Cum.
0	142,435	34.18	34.18
1	274,228	65.82	100.00
Total	416,663	100.00	

Tabela 4.14: *Distribuição da frequência da variável label2 para a base likes e posts*
Fonte: Elaboração própria

Onde, pode-se perceber que existem diferenças consideráveis entre o período em que determinado usuário apresentou sintomas depressivos e o período em que não apresentou os sintomas.
Também foi realizada uma investigação de como a variável *label4* se comporta ao longo do tempo.
As saídas obtidas encontram-se na figura 4.15

label2	label2 0	1	Total
0	100.00	0.00	100.00
1	0.00	100.00	100.00
Total	27.09	72.91	100.00

label2	label2 0	1	Total
0	100.00	0.00	100.00
1	0.00	100.00	100.00
Total	34.18	65.82	100.00

Tabela 4.15: *Comportamento de transição da variável label2 para a base likes e posts*
Fonte: Elaboração própria

Por meio dos resultados apresentados na figura 4.15, é possível verificar que existe considerável persistência do comportamento da variável *label4*, ou seja, analisando o comportamento diário de cada um dos usuários, independente de apresentarem ou não sintomas depressivos, eles tiveram o mesmo comportamento no dia seguinte, isso significa que a presença ou a ausência desses sintomas não interfere quantitativamente nas ações do usuário.
Elaboradas as análises preliminares, a próxima etapa, refere-se às práticas do modelo de regressão logística, segundo a teoria expressa no Capítulo 2.

O modelo de regressão que será apresentado leva em consideração a base de dados *likes* e *post* conjuntamente e a variável binária label2.

O coeficiente de correlação de Pearson é representado pela seguinte equação:

$$\rho = \frac{Cov\,(X_i, X_j)}{\sqrt{varX_i}\sqrt{varX_j}} \tag{4.4}$$

Os dados representados por "-" representam ausência de informação, ou seja, dados ausentes nessa base de dados, isso ocorre porque as bases não possuem exatamente todas as variáveis em comum. O comportamento apresentado pela correlação é relativamente semelhante para as variáveis entre as categorias, isto implica que não existe grande distinção entre a magnitude da correlação comparando as duas bases de dados.

Por exemplo, a variável independente *idade* apresenta correlação negativa para a base de dados *posts* e *likes* de – 42,2% e – 34,7%, respectivamente. Este coeficiente implica que quanto mais avançada a idade do usuário, menor é a probabilidade de apresentar sintomas depressivos, conforme mostra a tabela 4.12.

	label2	idade
label2	1.0000	
idade	-0.3471	1.0000

	label2	idade
label2	1.0000	
idade	-0.4216	1.0000

Figura 4.12: *Exemplo de correlação negativa para a variável idade*
Fonte: Elaboração própria

A Tabela 4.16 apresenta os resultados encontrados para o coeficiente de correlação de Pearson de cada variável com relação a variável *label2*, lembrando que a correlação ocorre entre duas variáveis e varia entre os valores de -1 (negativamente alinhada) e 1 (positivamente alinhada).

4.16: Coeficientes de correlação[12] para a variável *label2* (para a base de dados *posts* e *likes*)

[12] As lacunas representadas por -"correspondem ausência de informação.

Variáveis	Posts	Likes
apetite	52,6%	50,5%
choro	62,2%	61,9%
concentração	55,8%	54,3%
culpa	54,3%	50,6%
desvalorização	71,0%	65,8%
domingo	0,0%	1,6%
energia	61,1%	56,0%
fadiga	62,6%	59,4%
idade	-42,2%	-34,7%
indecisão	53,9%	51,6%
irritabilidade	53,6%	50,6%
prazer	62,7%	59,5%
quarta	-0,3%	-0,8%
sábado	0,6%	1,8%
segunda	0,1%	-0,3%
sexo	-7,3%	-10,3%
sexta	-0,2%	-0,3%
sono	44,5%	43,3%
qtd_message	2,7%	-
qtd_sent_neg	6,5%	-
qtd_sent_pos	5,6%	-
qtd_story	0,6%	-
t1	-1,2%	-2,1%
t2	-0,8%	-1,5%
t4	0,9%	2,5%
terça	-0,2%	-1,0%
total_dias	-17,6%	0,0%
total_friends	3,0%	-0,6%
-7,3%	24,3%	20,2%
total_posts	23,3%	19,5%
tarde	-	-1,7%
madrugada	-	4,2%
noite	-	-1,1%

Tabela 4.16: *Coeficiente de correlação de Pearson de cada variável com relação a variável label2.*
Fonte: Elaboração própria

O coeficiente de correlação estimado para a variável temporal que representam os dias da semana, apresenta que it sábado e *domingo* são os dias da semana com correlação positiva para os sintomas depressivos, ou seja, os usuários que apresentam maior probabilidade de apresentarem sintomas depressivos são os que que interagem mais no final de semana, quando comparados aos outros dias. Referente a variável sexo, o coeficiente de correlação para a base de dados *likes* é de -10,3 e para a base de dados *posts* é de -7,3%. Isto implica que aqueles que tem a menor probabilidade de apresentarem sintomas depressivos são os usuários do sexo masculino.

Quanto à variável que corresponde ao comportamento na rede social, it total_likes, tende a influenciar positivamente na presença de sintomas depressivos, este resultado é observado tanto para a base de dados *likes* quanto *posts*. Observa-se que sintomas depressivos, possuem uma relação positiva com a variável referente ao turno, especificamente no período da madrugada (das 0:00h às 6:00h) com coeficiente de 4,2%.

A Tabela 4.17, apresenta consolidados os coeficientes em logaritmo natural da razão de chance referentes ao modelo *probit pooled*.[13]. Este método assim, como as variáveis contidas em cada modelo foram especificadas pelo valor de máxima verossimilhança (*log likelihood*) e convergência do modelo especificado, onde o intuito é estimar os parâmetros de Z, para cada usuário i, por:

$$Z_i = a + \beta_1.idade_i + \beta_2.sexo_i + ... + \beta_k.X_{ki} \tag{4.5}$$

Onde, à partir da maximização do logaritmo da função de verossimilhança, apresenta a expressão:

[13] Dados não balanceados.

$$p_i = \Phi(Z_i) = \Phi(a + \beta_1.idade_i + \beta_2.sexo_i + ... + \beta_k.X_{ki}) \quad (4.6)$$

A ausência da variável no modelo, implica que a mesma não convergiu, entrando em uma área de não concavidade.

Tabela 4.17: Valor encontrado para os coeficientes estimados da razão de chance em ln, com o modelo *probit pooled* para a variável label2.

	Base de dados *posts*		Base de dados *likes*	
Variáveis	**Coef**	$p - valor\|\|$	**Coef**	$p - valor\|\|$
Cons	-0,86136	0,189	-2,83466	0
apetite	0,069377	0,612	0,092694	0,433
choro	0,130108	0,287	0,282684	0,013
concentração	0,165712	0,355	0,385251	0,018
culpa	0,523937	0,015	0,564185	0,011
desvalorização	0,854983	0	0,729218	0
domingo	-0,02592	0,3	-	-
energia	0,289662	0,108	0,20415	0,214
fadiga	0,269372	0,086	0,365913	0,015
idade	-0,03582	0,001	-0,02397	0,03
indecisão	-0,00523	0,968	0,177591	0,15
irritabilidade	0,32425	0,027	0,54526	0
prazer	0,62018	0,003	0,789866	0
quarta	-0,01891	0,237	-	-
sábado	-0,00031	0,989	-	-
segunda	-0,00874	0,596	-	-
sexo	0,391328	0,114	0,289207	0,237
sexta	-0,03081	0,046	-	-
sono	-0,18853	0,243	-0,34443	0,057
qtd_message	-0,00897	0,364	-	-
qtd_sent_neg	0,003734	0,768	-	-
qtd_sent_pos	0,001674	0,449	-	-
qtd_story	0,002888	0,793	-	-
t1	0,001892	0,914	0,005226	0,899
t2	-0,00042	0,981	-0,01676	0,636
t4	-0,01162	0,544	0,022246	0,502
terça	-0,02565	0,093	-	-
total_dias	-0,00036	0,085	-	-
total_friends	-0,00019	0,044	-0,0003	0,003
total_likes	-2,35E-06	0,988	0,000282	0,001
total_posts	0,000211	0,01	9,80E-06	0,953
tarde	-	-	-0,04397	0,129
madrugada	-	-	0,119448	0,006
noite	-	-	0,026854	0,358
Wald $chi^2(30)$ = 219,03			**Wald $chi^2(22)$ = 165,66**	
Prob > chi^2 = 0,0000			**Prob > chi^2 = 0,0000**	
Pseudo R^2 = 0,7284			**Pseudo R^2 = 0,7574**	
Log pseudolikelihood = -26412,132				

Tabela 4.17: *Coeficientes estimados para o modelo probit pooled.*
Fonte: Elaboração própria

O ajuste do modelo foi satisfatório ao explicar mais de 72% da ocorrência de sintomas depressivos para a base de dados *posts* e mais de 75% na base de dados *likes*, representados pelo valor R^2. A primeira coluna da Tabela 4.18, apresenta as variáveis que explicam a ocorrência de comportamento depressivo, tanto para a base de dados *posts* quanto para a base de dados *likes*. A segunda e quarta colunas apresentam os coeficientes em logaritmo natural da razão de chance para a base de dados *posts* e *likes*, respectivamente.

A terceira e quinta colunas apresentam o p-valor de cada coeficiente estimado, os valores inferiores a 0,01 são definidos com relação estatística igual ou superior a 99%, os coeficientes iguais ou inferiores a 0,05 representam relação estatística significativa igual ou superior a 95%, a mesma lógica é aplicada para os valores inferiores ou iguais a 0,1. Os demais valores, não apresentaram relação estatística válida.

Após os resultados estimados pelo modelo *probit*, é possível estimar a probabilidade condicional de ocorrência ou não de sintomas depressivos. Neste caso, a função de probabilidade é expressa da seguinte forma:

$$p_i = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-(a+\beta_1.X_{1i}+\beta_2.X_{2i}+...+\beta_k.X_{ki})}} \tag{4.7}$$

Com respaldo da expressão 4.7, pode-se apresentar um usuário que possui o valor 1 para as variáveis em sua escala para apetite, choro, *concentracao, culpa, desvalorizacao, qtd_message, qtd_sent_pos* e *qtd_story*. Além disso, este usuário é do sexo masculino, tem 25 anos de idade, está interagindo no turno da madrugada em um domingo no quarto trimestre do ano, possuindo um total de 98 postagens, 35 curtidas, 191 amigos e 1826 dias de atividade, as demais variáveis possuem peso zero, ou seja, são ausentes nas observações deste usuário. Desta forma, pode-se calcular a probabilidade condicional deste usuário substituindo os valores observados e seus respectivos coeficientes da Tabela 4.17 na equação 4.8.

Neste caso, foi substituído na equação, tanto os coeficientes estimados para a base de dados *posts* quanto para *likes*, a probabilidade encontrada é apresentada a seguir.

Probabilidade de sintomas depressivos (banco de dados *likes*)

$$p_i = \frac{1}{1+e^{-(-0,995)}} \cong 0,2698 \tag{4.8}$$

Probabilidade de sintomas depressivos (banco de dados *posts*)

$$p_i = \frac{1}{1+e^{-(-0,336)}} \cong 0,4166 \tag{4.9}$$

Desta forma, tem-se que para este usuário a probabilidade de apresentar sintomas depressivos é de 26,98% quando comparado com os dados apresentados para a base de dados *likes*, e de 41,66% para a base de dados *posts*, ou seja, um usuário que possui o mesmo perfil tem maior probabilidade de apresentar sintomas depressivos na interação com as postagens do que com as curtidas.

Uma das grandes vantagens da regressão probabilística é que cada coeficiente estimado fornece uma estimativa do logaritmo natural (ln) da razão de chance ajustado para todas as variáveis do modelo, permitindo a estimação direta da razão de chance por intermédio da exponenciação do coeficiente β, tem-se assim: e_i^{β}

Considerando os valores estimados na Tabela 4.17, e aplicando a exponencial, são apresentados na Tabela 4.18, as razões de chance. Observa-se para a base de dados *posts*, considerando os dados temporais dos dias da semana, que existe uma relação de 0,1% de chance a mais da presença dos sintomas depressivos nos usuários que interagem no primeiro trimestre, comparados aos usuários que interagem nos outros trimestres. Quanto aos dias da semana, não se apresentou uma razão de chance superior a 1 para os dias analisados.

Tabela 4.18: Razão de chances para a variável *label2* para a base de dados *posts* e *likes*

Variável	Post	Likes
Cons	0,422586	0,058739
apetite	1,071841	1,097126
choro	1,138951	1,326686
concentração	1,180233	1,469983
culpa	1,688662	1,758014
desvalorização	2,351334	2,073459
domingo	0,974418	-
energia	1,335976	1,226482
fadiga	1,309142	1,44183
idade	0,964815	0,976317
indecisão	0,994786	1,194336
irritabilidade	1,382994	1,725057
prazer	1,859262	2,2031
quarta	0,981269	-
sábado	0,999686	-
segunda	0,991293	-
sexo	1,478943	1,335368
sexta	0,969658	-
sono	0,828176	0,708622
qtd_message	0,991068	-
qtd_sent_neg	1,003741	-
qtd_sent_pos	1,001675	-
qtd_story	1,002892	-
t1	1,001894	1,005239
t2	0,999584	0,98338
t4	0,988447	1,022496
terça	0,974677	-
total_dias	0,999644	-
total_friends	0,999807	0,999704
total_likes	0,999998	1,000282
total_posts	1,000211	1,00001
tarde	-	0,95698
madrugada	-	1,126875
noite	-	1,027218

Tabela 4.18: *Razão de chances para a variável label2 para a base de dados posts e likes.*
Fonte: Elaboração própria

Os resultados encontrados para a base de dados *likes*, mostram que os usuários que interagem de madrugada, apresentam 12% a mais de sintomas depressivos do que os usuários que interagem nos demais horários do dia. Esta relação temporal também é observada para o quarto trimestre do ano, que apresenta sintomas depressivos em 2,2% a mais quando comparada com os demais períodos do ano. Em adição, os dados referentes a variável *desvalorizacao*, mostra que as pessoas que possuem alta pontuação para essa variável, possuem mais que o dobro de chance de apresentar sintomas depressivos comparado às pessoas que apresentam baixa pontuação, este é um dos principais fatores do comportamento do usuário relacionado aos sintomas depressivos.

Os dados para o total de amigos, total de curtidas e total de postagens, não apresentaram uma razão de chance (*odd ratio*) maior que 5% comparada com os casos que não possuem, desta forma, embora sejam significativos, não apresentam uma contribuição alta (maior que 5%) para explicar a presença de sintomas depressivos.

Além do teste de Wald, de significância conjunta das variáveis apresentadas na Tabela 4.19, também foram realizados alguns testes de Wald adicionais, apresentando os resultados obtidos para o teste de significância conjunta de Wald para cada variável estimada, este teste identifica se cada uma das variáveis que se encontra no modelo explica, em termos estatísticos, os valores observados para a variável label2 (com/sem sintomas depressivos). As lacunas representam que a variável não está presente no modelo estimado.

Tabela 4.19 - Resultados encontrados para o teste de Wald de significância conjunta

	Likes		Posts	
Label2	**Wald**	**Prob**	**Wald**	**Prob**
apetite	0,61	0,433	0,26	0,6117
choro	6,19	0,0129	1,13	0,2873
concentração	5,64	0,0175	0,85	0,3552
culpa	6,51	0,0107	5,94	0,0148
desvalorização	26,93	0	36,19	0
domingo	-	-	1,07	0,2999
energia	1,55	0,2136	2,59	0,1076
fadiga	5,95	0,0148	2,95	0,0859
idade	4,72	0,0297	11,11	0,0009
indecisão	2,07	0,1501	0	0,9683
irritabilidade	13,93	0,0002	4,92	0,0265
prazer	14,42	0,0001	9,07	0,0026
quarta	-	-	1,4	0,2374
sábado	-	-	0	0,9891
segunda	-	-	0,28	0,5963
sexo	1,4	0,2366	2,5	0,1136
sexta	-	-	3,99	0,0457
sono	3,63	0,0569	1,36	0,243
qtd_message	-	-	0,82	0,3641
qtd_sent_neg	-	-	0,09	0,7676
qtd_sent_pos	-	-	0,57	0,4486
qtd_story	-	-	0,07	0,793
t1	0,02	0,8989	0,01	0,9135
t2	0,22	0,6361	0	0,9808
t4	0,45	0,5019	0,37	0,544
terça	-	-	2,83	0,0927
total_dias	-	-	2,96	0,0852
total_friends	8,63	0,0033	4,04	0,0443
total_likes	10,78	0,001	0	0,9875
total_posts	0	0,9528	6,64	0,0099
tarde	2,31	0,1287	-	-
madrugada	7,68	0,0056	-	-
noite	0,84	0,3584	-	-

Tabela 4.19: *Resultados encontrados para o teste de Wald de significância conjunta*
Fonte: Elaboração própria

A primeira coluna da Tabela 4.19, apresenta os nomes das variáveis testadas, a segunda e quarta colunas, consolidam o valor encontrado para o teste de Wald, enquanto as colunas 3 e 5 vislumbram os resultados encontrados para o p-valor.

Os resultados obtidos para as variáveis com p-valor igual ou inferior a 5%, são significativos ao nível de confiança de 5% (p-valor igual a 0,05). Deste modo, há evidências de que cada uma das variáveis adicionadas ao modelo explica o comportamento da variável label2 (com/sem sintomas depressivos). Em adição, os valores de p-valor superiores a 5% comprovam que estas variáveis não explicam a existência ou não de sintomas depressivos, é o caso das variáveis temporais: *noite, tarde, t1, t2 e t4*, para a base de dados *likes*. Este resultado evidencia que a presença de sintomas depressivos nos usuários, não está relacionada à interação com o Facebook nestes períodos de tempo, e vale observar que a variável temporal *madrugada*, que reflete a interação com o Facebook das 00:00h às 6:00h, é a única variável temporal significativa, ou seja, mostra a relação com sintomas depressivos nos usuários.

Ainda na base de dados *likes*, a presença ou não de sintomas depressivos nos usuários não são justificáveis pelas variáveis *indecisão* e *apetite*.

Na base de dados *posts*, observa-se que a presença ou ausência de sintomas depressivos nos usuários, não são estão sendo explicadas pelas variáveis temporais referentes aos trimestres do ano, estes valores são reforçados através da tabela de razão de chance que apresenta uma pequena diferença, inferior a 5%, entre estar ou não interagindo com o Facebook no período de tempo analisado.

Observa-se ainda que a ocorrência de casos de sintomas depressivos não é explicado pelo total de likes, na base de dados *posts*, assim como as variáveis *qtd_message, qtd_sent_neg, qtd_sent_pos e qtd_story*.

Estes resultados, também são evidenciados na tabela de razão de chances onde não apresentam importância relativa de causa ou não causa da presença do comportamento analisado.

Por fim, a existência de sintomas depressivos na base de dados *posts* foram explicados pelas variáveis: *culpa, desvalorização, fadiga e energia* (significativo a 90%), também as *idade, irritabilidade e prazer*, e referentes ao período temporal foram asa variáveis *sexta-feira e terça-feira, total de dias, amigos e posts.*

De acordo com os testes executados e expostos anteriormente nesse capítulo, de uma maneira geral, o modelo de regressão logítica multinomial, foi considerado superior ao modelo de regressão logítica binária. Para melhorar o entendimento dessa afirmação, uma análise contrastando os dois modelos com as bases de dados likes e posts foi realizada e será descrita à seguir. Com o objetivo de contrastar o modelo proposto subdividido em multinomial e binário, alguns testes adicionais amparados por medidas estatísticas, foram realizados.

Enquanto algumas medidas calculadas são fundadas em valores retornados pelo comando de estimativa, para algumas medidas é necessário calcular estatísticas adicionais da amostra de estimativa, segundo Scott (2001).

4.8 Análise de contraste dos modelos multinomial e binário

Vários testes foram realizados na busca de um modelo com melhor poder preditivo, deste modo, os modelos de regressão logística multinomial e binária apresentado nas seções 4.6 e 4.7, respectivamente, foram selecionados e considerados, dentre aqueles que foram testados, os mais adequados para explicar o fenômeno em estudo. Alguns resultados de testes podem ser repetidas nessa etapa apenas por questões de inteligibilidade.

Segundo Fávero e Belfiore [2017], após um modelo de regressão logística ter sido ajustado, um teste global de adequação do modelo resultante deve ser realizado.

Em conformidade com Allison [2014], para comprovar que o modelo se ajusta aos dados, duas abordagens podem ser consideradas, a primeira delas são as medidas de poder preditivo, que incluem R2 de McFadden, a área sob a curva ROC e várias correlações de ordem de classificação, a segunda abordahem refere-se aos testes de ajuste da qualidade do modelo *goodness of fit* (GOF), onde podem ser utilizados o *deviance*, o qui-quadrado de Pearson ou o teste de Hosmer-Lemeshow, estes visam auxiliar na decisão da correta especificação do modelo.

Ainda segundo Zhang *et al.* [2017], o passo final de um modelo de regressão logística é verificar a qualidade do ajuste do modelo, afirmando que o GOF usa uma estatística resumida para avaliação do ajuste do modelo, incluindo a estatística qui-quadrado de Pearson, *deviance*, soma dos quadrados e Testes de Lemeshow. Estas estatísticas medem a diferença entre os valores observados e ajustados.

Portanto, testes para evidenciar que o modelo logístico multinomial se ajusta melhor aos dados Vivamente é o que será realizado na presente etapa.

4.8.1 Contraste considerando a base de dados Likes

Primeiramente os modelos com a base de dados *likes* foram avaliados.

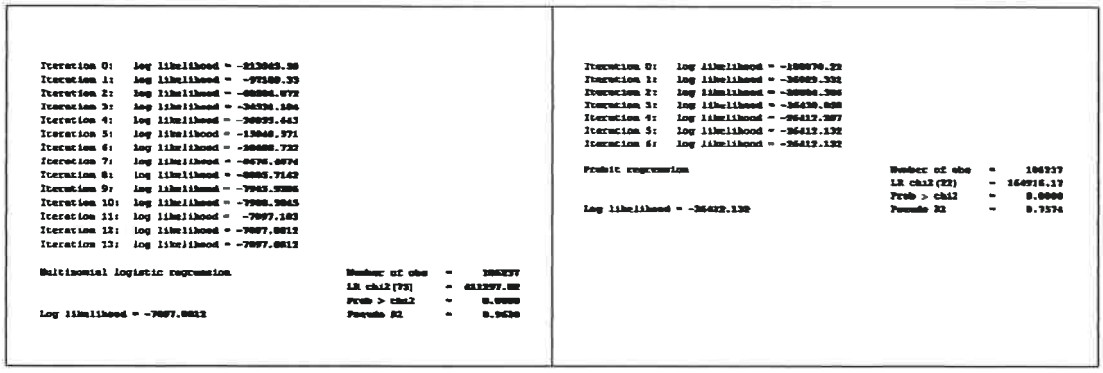


Tabela 4.20: *Análise de regressão logística multinomial e binária*
Fonte: Elaboração própria

Observa-se na figura 4.20, para o modelo **multinomial**, uma listagem das probabilidades do log em cada iteração do modelo de regressão. A primeira iteração (denominada iteração 0) é a probabilidade do log

do modelo "nulo"ou "vazio", na próxima iteração, os preditores são incluídos no modelo, o primeiro valor dessa iteração foi de -213.545,59, nas iterações seguintes a probabilidade de log aumenta a cada iteração até convergir, nesse ponto a iteração é interrompida e os resultados são exibidos, para esse modelo foram 13 iterações e o valor final alcançado foi de -7.897,08, a categoria 1 foi a mais freqüente e portanto definida como o grupo de referência neste modelo. Em contrapartida, para o modelo **binário** foram 6 iterações, onde seus valores iniciaram em -108.870,22 e tendo alcançado a convergência em -26.412,13. Portanto as probabilidades de log dos modelos ajustados foram de -7.897,08 e -26.412,13, para os modelos multinomial e binário, respectivamente. Obedecendo a função objetivo apresentada na equação 2.53 para o modelo multinomial e a equação 2.25 para o modelo binário, ou seja, que o valor do somatório do logaritmo da função de verossimilhança seja o máximo possível, nesse sentido afirma-se que o modelo logístico multinomial apresentou proporcionalmente o menor valor, sendo portanto considerado com a melhor estimação.

No que tange as variáveis preditivas analisadas individualmente, alguns exemplos podem ser ilustrados, a exemplo da variável *prazer*, por exemplo, teve uma estimativa de 7,009547, sendo que esta é a estimativa logística **multinomial** para um aumento de uma unidade na pontuação do sintoma relacionado a perda de prazer para a categoria leve em relação à categoria moderado, dado que as outras variáveis no modelo são mantidas constantes. Se um usuário aumentasse sua pontuação na perda de prazer em um ponto, seria esperado que a probabilidade multinomial de preferência de leve para moderado aumentasse em 7 unidades, mantendo todas as outras variáveis constantes do modelo. O mesmo acontece para a variável *suicida* e *choro*, com valores de 6,670905 e 5,389144, respectivamente. Relacionada a estimativa logística **binária** a variável *prazer* e *choro*, obtiveram os valores de 0,7898655 e 0,2826839, respectivamente, já a variável *suicida* não convergiu para esse modelo. Nesse aspecto pode-se reconhecer que algumas das variáveis preditivas importantes para a significância de um modelo, não poderiam deixar de ser consideradas, bem como a sua diferença entre uma categoria e a sua próxima categoria, por exemplo de leve para moderado (considerando as 4 categorias), aspecto esse que no modelo **binário**, devido a sua nomenclatura, não foram considerados.

O teste qui-quadrado da razão de verossimilhança (LR chi2) foi de 411.297 e 164.916, indicando que pelo menos um dos coeficientes de regressão dos preditores não é igual a zero. No teste Prob> chi2, sendo essa a probabilidade de obter a estatística qui-quadrado dado que a hipótese nula é verdadeira, ambos os modelos demonstram ser estatisticamente significativos pois seus valores de p foram menores que 0,000.

O teste pseudo quadrado de McFadden ($PseudoR^2$), indica que aproximadamente 96% e 76% da variação da variável dependente pode ser explicada pelas variáveis independentes dos modelos **multinomial** e **binário**, respectivamente, essa saída mostra que o modelo para o modelo multinomial as variáveis explicam melhor o modelo.

De maneira geral pode-se afirmar que o modelo logístico **multinomial** obteve resultados mais significativos em contraste ao modelo logístico **binário**.

Os testes apresentados na sequência, à partir da figura 4.21, são algumas medidas extra de ajuste dos modelos.

Log-likelihood	
Model	-7893.640
Intercept-only	-213545.593
Chi-square	
Deviance (df=186156)	15787.281
LR (df=78)	411303.905
p-value	0.000
R2	
McFadden	0.963
McFadden (adjusted)	0.963
Cox-Snell/ML	0.890
Cragg-Uhler/Nagelkerke	0.990
Count	0.991
Count (adjusted)	0.981
IC	
AIC	15949.281
AIC divided by N	0.086
BIC (df=81)	16778.198

Log-likelihood	
Model	-26412.132
Intercept-only	-108870.216
Chi-square	
Deviance (df=186214)	52824.264
LR (df=22)	164916.168
p-value	0.000
R2	
McFadden	0.757
McFadden (adjusted)	0.757
McKelvey & Zavoina	0.914
Cox-Snell/ML	0.587
Cragg-Uhler/Nagelkerke	0.852
Efron	0.774
Tjur's D	0.774
Count	0.936
Count (adjusted)	0.763
IC	
AIC	52870.264
AIC divided by N	0.284
BIC (df=23)	53103.364
Variance of	
e	1.000
y-star	11.691

Tabela 4.21: *Medidas de ajuste para o modelo multinomial e binário*
Fonte: Elaboração própria

O primeiro teste exibido foi o de medida do log de verossimilhança de todos os parâmetros e na sequência a probabilidade de log após a convergência também foi listada, onde obtiveram o valor de 213.545,59 e 108.870,21 para os modelos **multinomial** e **binário**, respectivamente.

Após, os testes qui-quadrado foram apresentados, referindo-se ao *Deviance*(df=186156), com valor de 15.787,281 para o modelo **multinomial** e Deviance (df=186214) com valor de 52.824,264 para o modelo **binário**. Logo na sequência, apresenta-se o teste *Likelihood Ratio* (LR), que para o modelo **multinomial** foi reportado com LR (78) = 411.303,905, já para o modelo **binário**, o LR foi reportado com LR(22) = 164.916,168, com os graus de liberdade, 78 e 22, respectivamente.

Os testes que se referem ao teste R^2, que reportam o coeficiente padrão de determinação, analisados foram os testes de: McFadden com valor de 0,963, o McFadden ajustado com 0,963, o Cox-Snell/ML a 0,890 Cragg-Uhler/Nagelkerke quantificando 0,990 o Count pontuando 0,991, e por fim o Count ajustado com valor de 0,981, todos esses valores de R^2 são referentes ao modelo **multinomial**. Para o modelo **binário**, apresentam-se os seguintes valores: o McFadden pontuou 0,757 o McFadden adjustado foi igual a 0,757 o Cox-Snell/ML obteve 0,587 o Cragg-Uhler/Nagelkerke 0,852 o Count 0,936 e por último o Count ajustado com valor de 0,763. Considerando todos os testes R^2 apresentados para os dois modelos, pode-se dizer que o modelo **binário** teve em cada um dos testes, seus valores abaixo dos valores para os mesmos testes do modelo **multinomial**, indicando assim, menor ajuste aos seus dados.

No que tange aos critérios de informação *Bayesian Information Criterion* (BIC) e *Akaike information criterion* (AIC), para o modelo **multinomial** o critério AIC apresenta um valor de 15.949,28 contra o valor de 52.870,26 apresentado para o modelo **binário**, já para o critério BIC os valores são de 16.770,19 e de 53.103,36 para os modelos **multinomial** e **binário**, respectivamente, considerando que o modelo com o menor AIC ou BIC é considerado o modelo de melhor ajuste, afirma-se que o modelo **multinomial**, obteve melhor ajuste perante esses dois critérios.

Na sequência serão apresentadas medidas GOF para o modelo. As tabelas 4.22 e 4.23, apresentam os resultados em porcentagens das matrizes de confusão, primeiramente para o modelo logístico multinomial e na sequência para o modelo logístico binário.

Nível de depressão	Mínimo	Leve	Moderado	Grave
Mínimo	99,77%	0,23%	0%	0%
Leve	1,10%	96,09%	2,81%	0%
Moderado	0%	4,26%	95,17%	0,57%
Grave	0%	0%	1,13%	98,87%
Precisão	98,42%			

Tabela 4.22: *Matriz de confusão multinomial*
Fonte: Elaboração própria

Depressão	Sim	Não
Sim	95,28%	4,72%
Não	11,13%	88,87%
Precisão	93,58%	

Tabela 4.23: *Matriz de confusão binária*
Fonte: Elaboração própria

Observa-se por meio da tabela 4.22, que o modelo logístico **multinomial**, apresenta uma precisão geral de 98,42 %. Considerando que a precisão pode não ser uma métrica confiável para o desempenho real de um classificador se o conjunto de dados estiver desequilibrado, como é o caso dos dados dos modelos, apresenta-se a seguir a matriz de confusão para o modelo **multinomial**, sendo que para a classe que representa a categoria *mínimo* foi de 99,77%, para a *leve* foi 96,09%, a *moderado* alcançou 95,17% e para a *grave* foi de 98,87%.

A tabela 4.23, apresenta, por outro lado, o modelo logístico **binário** que de modo geral prevê 93,58% das observações corretamente, tendo classificado como verdadeiro positivo 95,28% dos usuários, ou seja, 95,28% dos usuários que possuíam sintomas depressivos foram classificados na classe correta (com depressão), e 88,87% foram verdadeiros negativos, significando que 88,87% dos usuários com sintomas depressivos foram classificados corretamente, já 11,13% foram falsos positivos, quer dizer que foram classificados na classe de possuidores de sintomas depressivos mas na verdade não apresentavam os sintomas, e 11,13% dos usuários foram denominados como falsos negativos, indicando que cairam na classe dos não depressivos, sendo que apresentavam os sintomas.

Considerando as matrizes de correlação para os modelos logístico e multinomial para a base de dados *likes*, conclui-se que os modelos obtiveram uma ótima classificação, embora o modelo multinomial mostrou-se superior.

4.8.2 Contraste considerando a base de dados Posts

Nessa seção os modelos com a base de dados *posts* foram avaliados.

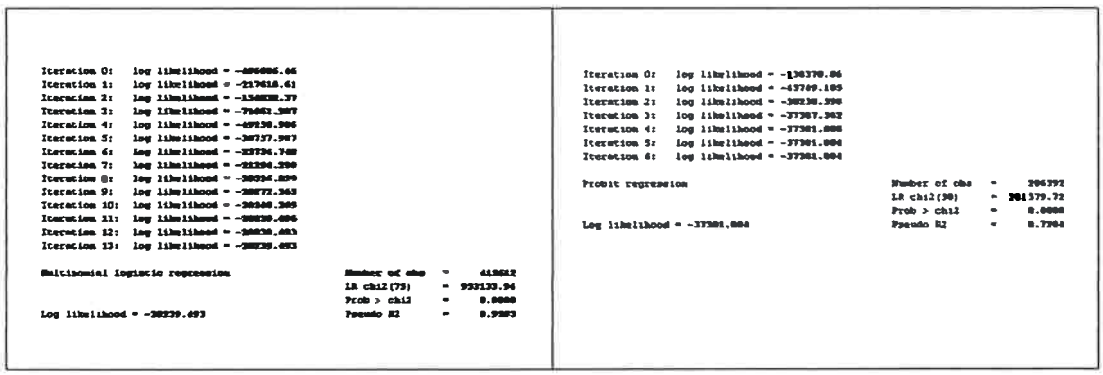


Tabela 4.24: *Análise de regressão logística multinomial e binária*
Fonte: Elaboração própria

Sendo a máxima verossimilhança um procedimento iterativo, utilizado pela regressão logística, a primeira informação a ser observada é uma listagem das probabilidades do log em cada iteração, sendo a primeira

iteração a probabilidade do log do modelo nulo, na próxima iteração, os preditores são incluídos no modelo. A tabela 4.24, que para o modelo **multinomial**, o primeiro valor foi -496.806,46, aumentando nas iterações seguintes até convergir, momento em que a iteração foi interrompida e os resultados foram exibidos, para esse modelo foram 13 iterações e o valor final alcançado foi de -20.239,49, sendo a categoria 1 a mais freqüente, foi definida como sendo a referência neste modelo. Em contrapartida, para o modelo **binário** foram 6 iterações onde seus valores iniciaram em -138.370,86 e tendo alcançado a convergência em -37.581. Portanto as probabilidades de log dos modelos ajustados foram de -20.239,49 e -37.581, para os modelos multinomial e binário, respectivamente. Atendendo a função objetivo apresentada na equação 2.53 para o modelo multinomial e a equação 2.25 para o modelo binário, que assegura que o valor do somatório do logaritmo da função de verossimilhança seja o máximo possível, nesse sentido afirma-se que o modelo logístico multinomial apresentou proporcionalmente o menor valor, sendo portanto considerado com a melhor estimação.

O teste qui-quadrado da razão de verossimilhança (LR chi2) foi de 953.133,94 e 201.579,72, indicando que pelo menos um dos coeficientes de regressão dos preditores não é igual a zero. No teste Prob> chi2 sendo esta a probabilidade de obter a estatística qui-quadrado, dado que a hipótese nula é verdadeira, e em ambos os modelos pode-se afirmar que o modelo é estatisticamente significativo porque o valor de p é menor que 0,000.

O teste pseudo quadrado de McFadden ($PseudoR^2$), indica que aproximadamente 96% e 73% da variação da variável dependente pode ser explicada pelas variáveis independentes dos modelos **multinomial** e **binário**, respectivamente.

De maneira geral pode-se afirmar que o modelo logístico multinomial obteve resultados superiores em contraste ao modelo logístico binário.

Os testes apresentados na sequência, à partir da tabela 4.25, são algumas medidas extra de ajuste dos modelos.

Log-likelihood	
Model	-20239.493
Intercept-only	-496806.468
Chi-square	
Deviance (df=415534)	40478.985
LR (df=75)	953133.935
p-value	0.000
R2	
McFadden	0.959
McFadden (adjusted)	0.959
Cox-Snell/ML	0.899
Cragg-Uhler/Nagelkerke	0.998
Count	0.988
Count (adjusted)	0.976
IC	
AIC	40634.985
AIC divided by N	0.098
BIC (df=78)	41488.111

Log-likelihood	
Model	-37581.004
Intercept-only	-138370.862
Chi-square	
Deviance (df=206561)	75162.008
LR (df=30)	201579.717
p-value	0.000
R2	
McFadden	0.728
McFadden (adjusted)	0.728
McKelvey & Zavoina	0.893
Cox-Snell/ML	0.623
Cragg-Uhler/Nagelkerke	0.844
Efron	0.762
Tjur's D	0.763
Count	0.911
Count (adjusted)	0.772
IC	
AIC	75224.008
AIC divided by N	0.364
BIC (df=31)	75541.402
Variance of	
e	1.000
y-star	9.326

Tabela 4.25: *Medidas de ajuste para a o modelo multinomial e binário*
Fonte: Elaboração própria

O primeiro teste exibido foi o de medida do log de verossimilhança de todos os parâmetros e na sequência a probabilidade de log após a convergência também foi listada onde obteve o valor de 496.806,46 e 138.370,21

para os modelos **multinomial** e **binário**, respectivamente.

Na sequência, os testes qui-quadrado foram apresentados, referindo-se ao *Deviance*(df=415534), com valor de 40.478,98 para o modelo **multinomial** e Deviance (df=206561) com valor de 75.162 para o modelo **binário**. Logo na sequência, apresenta-se o teste *Likelihood Ratio* (LR), que para o modelo **multinomial** foi reportado com LR (75) = 953.133,93, já para o modelo **binário**, o LR foi reportado com LR chi2 (22) = 164.916,17, onde os graus de liberdade, 75 e 22, são o número de parâmetros restritos, respectivamente.

Os testes que se referse ao teste R^2, que reportam o coeficiente padrão de determinação podem ser definidos de várias formas, e para esse estudo foram selecionados os de McFadden com valor de 0,959, o McFadden ajustado com 0,959, o Cox-Snell/ML a 0,899 Cragg-Uhler/Nagelkerke quantificando 0,990 o Count pontuando 0,988, e por fim o Count ajustado com valor de 0,976, todos esses valores de R^2 são referentes ao modelo **multinomial**. Para o modelo **binário**, apresentam-se os seguintes valores: o McFadden pontuou 0,728 o McFadden adjustado foi igual a 0,728 o Cox-Snell/ML obteve 0,623 o Cragg-Uhler/Nagelkerke 0,844 o Count 0,911 e por último o Count ajustado com valor de 0,772. O teste Cox-Snell/ML, quase nunca atinge o valor máximo de 1, geralmente está perto de 0,75, mas o teste de Nagelkerke é praticamente um ajuste de Cox-Snell/ML para que atinja o valor máximo de 1, sendo assim será maior do que o valor de Cox-Snell/ML. Considerando todos os testes R^2 apresentados para os dois modelos, pode-se dizer que o modelo **binário** teve em cada um dos testes, seus valores abaixo dos valores para os mesmos testes do modelo **multinomial**, indicando assim, menor ajuste aos seus dados.

No que tange aos critérios de informação *Bayesian Information Criterion* (BIC) e *Akaike information criterion* (AIC), para o modelo **multinomial** o critério AIC apresenta um valor de 40.634,98 contra o valor de 75.224 apresentado para o modelo **binário**, já para o critério BIC os valores são de 41.488,11 e de 75.541,40 para os modelos **multinomial** e **binário**, respectivamente, considerando que o modelo com o menor AIC ou BIC é considerado o modelo de melhor ajuste, afirma-se que o modelo **multinomial**, obteve melhor ajuste perante esses dois critérios.

Na sequência serão apresentadas medidas de teste da qualidade do ajuste do modelo de GOF.

As tabelas 4.26 e 4.27, apresentam os resultados em porcentagens das matrizes de confusão, primeiramente para o modelo logístico multinomial e na sequência para o modelo logístico binário.

Nível de depressão	Mínimo	Leve	Moderado	Grave
Mínimo	99,75%	0,25%	0%	0%
Leve	0,35%	97,52%	2,13%	0%
Moderado	0%	4,26%	95,17%	0,57%
Grave	0%	2,17%	4,8%	92,97%
Precisão	98,38%			

Tabela 4.26: *Matriz de confusão multinomial*
Fonte: Elaboração própria

Depressão	Sim	Não
Sim	93,56%	6,44%
Não	12,58%	87,42%
Precisão	91,07%	

Tabela 4.27: *Matriz de confusão binária*
Fonte: Elaboração própria

Observa-se por meio da tabela 4.26, que o modelo logístico **multinomial**, apresenta uma precisão geral de 98,38 % e a seguir a matriz de confusão para o modelo **multinomial**, será apresentada sendo que para a classe que representa a categoria *mínimo* foi de 99,75%, para a *leve* foi 97,52%, a *moderado* alcançou 95,17% e para a *grave* foi de 92,97%.

A tabela 4.27, apresenta, por outro lado, o modelo logístico **binário** que de modo geral prevê 91,07% das observações corretamente, tendo classificado como verdadeiro positivo 93,56% dos usuários, ou seja, 93,56% dos usuários que possuíam sintomas depressivos foram classificados na classe correta (com depressão), e 87,42% foram verdadeiros negativos, significando que 87,42% dos usuários com sintomas depressivos foram classificados corretamente, já 6,44% foram falsos positivos, quer dizer que foram classificados na classe de possuidores de sintomas depressivos mas na verdade não apresentavam os sintomas, e 12,58% dos usuários

foram denominados como falsos negativos, indicando que cairam na classe dos não depressivos, sendo que apresentavam os sintomas.

Considerando as matrizes de correlação para os modelos logístico e multinomial para a base de dados *posts*, conclui-se que os modelos obtiveram uma classificação bastante positiva, embora o modelo multinomial tenha sido superior.

As saídas completas dos modelos no que se refere aos testes apresentados em toda essa seção estão no github[14].

Os testes aplicados nessa fase foram escolhidos em detrimento a outros pela facilidade de aplicação e interpretação, bem como pela sua frequente utilização em objetivos semelhantes possibilitando dessa forma um contraste mais consistente.

A seguir apresenta-se alguns exemplos de estudos onde testes GOF foram utilizados para a avaliação da qualidade do ajuste de modelos logísticos.

Allison [2014], afirma que tanto o *deviance* quanto o qui-quadrado de Pearson têm boas propriedades quando o número esperado de eventos e o número esperado de não eventos para cada perfil é de pelo menos 5. O que se enquadra nesse estudo, onde a quantidade de eventos e não eventos é bem superior a esse número.

O estudo de Scafato *et al.* [2012], tendo como propóstito analisar associações entre aspectos psicológicos, sociodemográficos e funcionais sobre o risco de isolamento social, mortalidade e re-hospitalização em idosos, em um estudo longitudinal com idosos italianos, assim como no presente estudo utilizou análise de regressão logística múltipla, onde utilizou o χ^2 como teste do GOF no modelo completo com todas as variáveis independentes contra um modelo somente constante tendo sido estatisticamente confiável com valores $\chi^2 = 102{,}86$, $p < 0{,}001$, indicando que as variáveis distinguiram de forma confiável entre risco de isolamento e não risco.

Santini *et al.* [2016], igualmente em seu estudo longitudinal com idosos irlandeses sobre o envelhecimento utilizou-se de regressão linear multivariada para avaliar as associações por meio de comparações das variáveis categóricas entre os grupos com testes qui-quadrado.

E para fechar, o estudo de Kayode *et al.* [2012] não poderia deixar de ser mencionado, seu objetivo foi desenvolver um modelo preditivo para identificar fatores de risco materno, infantil, familiar e outros associados à mortalidade em crianças abaixo dos 5 anos de idade, para alcançar seus objetivos utilizou a regressão logística univariada para examinar a associação entre as variáveis explicativas e o desfecho dependente, tendo também usufruido de testes bastante reconhecidos e utilizados como o teste da Razão de Verossimilhança (teste LHR) foi usado para testar a qualidade do ajuste do modelo.

Diante dos testes executados aliados aos estudos na área, no que diz respeito aos modelos de regressão logística considerando os aspectos avaliativos do poder preditivo do modelo, da adequação do ajuste do modelo e do teste de qualidade do modelo, pode-se afirmar que esse conjunto de recursos obteve resultados bastante positivos na busca pelos objetivos inerentes ao estudo apresentado nessa pesquisa.

4.8.3 Sintese dos processos

Essa seção traz um apanhado geral dos processos executados nesse estudo, a figura 4.13 apresenta esse fluxo que será descrito na sequência.

[14] https://github.com/cmaricy/Vivamente http://vivamente.herokuapp.com/

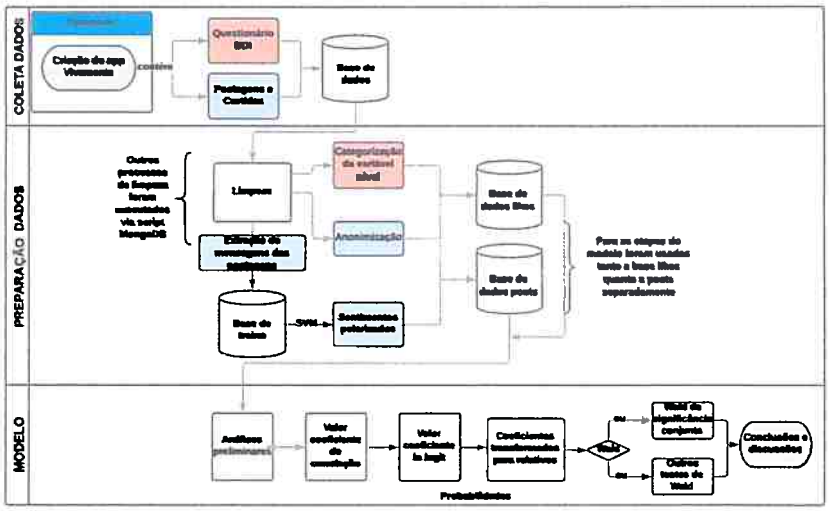


Figura 4.13: *Fluxo resumido dos processos*
Fonte: Elaboração Própria

Esta investigação aprovada pelo comitê de ética buscou contribuir de maneira interdisciplinar na área da ciência da computação, com abrangência na área da saúde e matemática. O estudo de mineração de dados em redes sociais e análise de sentimentos, com vistas a predição de traços depressivos exigiu uma exploração minuciosa de suas técnicas.

Para tanto, dados públicos de duas bases de dados foram utilizados, uma correspondente à extração de dados abertos por meio de uma ferramenta chamada Netvizz e outra já extraída, referente a dados oriundos do Twitter; ambas em formato de única cross section.

Algumas análises foram realizadas como pré-teste onde demonstraram a potencialidade de técnicas de mineração de dados na busca por informações relevantes, porém nenhuma demonstrou potencial para identificação de qualquer traço depressivo; tanto pela falta de dados em grande quantidade, quanto pela inexistência de dados específicos que pudessem revelar o nível de depressão de cada usuário.

Diante deste cenário, busca-se desenvolver uma ferramenta específica para coleta de dados dos usuários do Facebook. Após estudo sobre funcionamento do acesso aos dados do Facebook optou-se pelo acesso via Scrips Javascript, feitos os devidos testes via Script, observou-se que essa não seria uma boa alternativa pois a mesma apenas coletava dados do lado cliente, sem segurança e possibilitando somente acesso aos dados públicos do usuário, mas como a necessidade era pela captação de informações referente às postagens e curtidas, ou seja, dados restritos. Frente à exigência do Facebook, se fez necessária a criação de um aplicativo (app) do lado do servidor, onde o aplicativo permaneceria seguro.

A priori, optou-se pela construção do app utilizando uma API/SDK chamada Spring Social, realizando interface com a Graph API do Facebook, com suporte do Eclipse como ambiente de desenvolvimento, o GlassFish como servidor de aplicativos, o Apache Maven como gerenciador de dependências, etc, todos alinhados às tecnologias JEE (Java Enterprise Edition). Sendo que a escolha se deu pela familiaridade com a linguagem de programação Java e tecnologias relacionadas. Porém, no decorrer do desenvolvimento do app, houve a necessidade de declinar da decisão pela utilização dessas tecnologias, por dois motivos: o primeiro pela morosidade em realizar o deploy do app e em rodar parte do aplicativo já desenvolvido sendo testado, e o segundo pela falta de material de apoio para suportar tal desenvolvimento.

Neste sentido, optou-se por uma gama de ferramentas mais leves e tecnologia mais acessível no quesito, material de apoio, iniciando-se uma nova versão do app de coleta denominado Vivamente. Aplicativo este desenvolvido sob tecnologias NodeJS, com auxílio do Sublime Text para edição do código fonte, o NodeJS, como o interpretador de código JavaScript do lado do servidor; Express, como o framework para aplicativos Node.js; Passport, como middleware para autenticação da aplicação; AngularJS, para gerenciamento do questionário e MongoDB como banco de dados orientado a documentos.

O app Vivamente é composto de duas partes, a primeira refere-se ao Facebook canvas, que precisou ser criado e configurado dentro do Facebook, o canvas contém o ID do usuário, uma chave secreta fornecidos pelo Facebook e outras informações que possibititaram a aprovação do app sob a política do Facebook, autorizando assim a coleta dos dados dos usuários relativos às postagens e curtidas. A segunda é referente a um projeto web; desenvolvido e embutido no canvas, o qual contém um questionário denominado Inventário de Depressão de Beck II (BDI-II), com 21 perguntas relacionadas a investigação de traços de comportamento depressivo.

Inventário este selecionado após ampla pesquisa, o que permitiu perceber sua maturidade e larga aplicabilidade em pesquisas científicas, sendo uma ferramenta de escalas investigativas de sintomas depressivos com resultados positivos. Para que fosse possível fazer uso do Inventário de Depressão de Beck (BDI-II), a editora casa do Psicólogo Pearson, concedeu uma autorização para a utilização restrita a esse estudo tendo a psicóloga Sabrina Zaffari Farias, responsável pelo projeto no tocante a construção, manipulação e aplicação do questionário, garantindo que todas as normas fossem cumpridas. Ressalta-se que o estudo ainda foi amparado por mais dois psicólogos com vistas ao correto avanço da pesquisa.

O *app* Vivamente foi primeiramente armazenado no servidor *Google App Engine* (GAE), mas por questões de incomcompatibilidade e alto custo, passou a ser operado pelo servidor Heroku, e os dados armazenados no servidor Atlas MongoDB, ambas versões pagas para garantir a segurança, disponibilidade e confiabilidade. Terminada a implantação do *app* e aprovado pela revisão do Facebook, iniciou o processo de divulgação para a obtenção de voluntários.

A divulgação e coleta ocorreu inicialmente durante dois meses, mas como o total de dados de registros de usuários (296) esteve abaixo do necessário e desbalanceado o período de coleta extendeu-se por mais 4 meses, obtendo 692 registros válidos, embora ainda desbalanceado, mas considerado uma quantidade adequada para obtenção de resultados mais significativos, iniciou a fase de preparação dos dados.

Na preparação dos dados, foram utilizadas técnicas de exclusão, anonimização e agrupamento dos dados formando uma base de dados agrupada, porém seus resultados não foram reveladores, seguiu-se então para a categorização dos dados referentes ao nível de sintomas depressivos da variável *nivel* gerando uma base de dados orientada pelas datas das ações dos usuários, como as postagens e curtidas ocorrerem em datas distintas, duas bases de dados longitudinais foram geradas, denominadas *Posts* e *Likes*. Na sequência os dados dos textos das postagens puderam ser extraídos e classificados em sentimentos positivos e negativos, para conseguir uma boa classificação, uma base de dados de treinamento precisou ser criada, nesse processo de classificação, a técnica SVM obteve melhor precisão.

A análise dos dados utilizou-se em linhas gerais da Regressão Logística, nas análises preliminares efetuaram-se testes para verificação da não violação dos pressupostos de multicolinearidade, heterocedasticidade e a auto-correlação, resultando na não violação desses 3 pressupostos, para a comparação de médias nas amostras *com* e *sem* sintomas depressivos, executou-se o teste-t, demonstrando que as amostras não são iguais, o teste de distribuição de frequência da variável *label4* mostrou haver diferenças consideráveis entre os períodos que se apresentam os sintomas depressivos e os que não os apresentam, por meio do teste de comportamento da variável *label4* ao longo do tempo verificou-se sua persistência. Os testes para o coeficiente de correlação de cada variável com relação a variável *label4* e uma variável (*dummie*) demonstrou que o comportamento é semelhante para as variáveis ao longo das categorias.

No teste de estimação dos coeficientes da *ln* da razão de chance *odd ratio*, para o modelo logístico multinomial, os resultados obtidos mostram que a maioria das variáveis são significativas ao nível de confiança de 95%. No teste de probabilidade condicional de ocorrência dos níveis de sintomas depressivos em relação às variáveis em análise, o teste de significância conjunta de Wald mostrou que as variáveis explicam bem o modelo, já os testes adicionais de Wald apresentaram alguns valores negativos entre as categorias de sintomas depressivos.

Os testes para o Modelo de Regressão Logística Multinomial considerando a base de dados *posts*, foram os mesmo aplicados à base de dados *likes* e com os mesmo objetivos, diferindo apenas em alguns dos resultados.

Para os testes no Modelo de Regressão Logística Binária, as bases de dados *likes* e *posts* foram consideradas conjuntamente, a variável *label2* representou as duas categorias (com e sem depressão) e os testes foram basicamente os mesmos realizados para o modelo logístico multinomial.

Para finalizar, testes de medida do log de verossimilhança, probabilidade de log após a convergência, quiquadrado Deviance e Likelihood Ratio (LR), R2 de McFadden e Cox-Snell/ML, Cragg-Uhler/Nagelkerke e Count ajustado, critérios de informação BIC e AIC e matrizes de confusão foram executados, para a base de dados *posts* e *likes*, onde o modelo logístico multinomial mostrou-se superior em todos os testes realizados.

Considerando as interpretações dos coeficientes em termos de log-odds multinomial (logits) e as interpretações dos coeficientes em termos de razões de risco relativo para as variáveis prazer, fadiga, choro, interesse, desvalorização, fracasso, critica, estima, concentração, irritabilidade, indecisão, tristeza, pessimismo, suicida, apetite, sono, culpa, energia, t1, t2, t4, total_friends, total_likes, total_posts, sexo e idade, do modelo logístico multinomial, destacam-se algumas afirmações.

Afirma-se que todas as variáveis são significativas ao nível de confiança de 95%.

O resultado da variável do número de amigos mostra que o acréscimo no número de amigos desmotiva os usuários com maior nível de comportamento depressivo a acessarem a rede social.

A comparação com a base de dados likes mostra que os valores estimados para o primeiro e o segundo trimestres, para a categoria moderado, não são significativos para a base de dados posts, ao passo que eram

significativos para a base de dados likes, de modo semelhante, para o quarto trimestre, a categoria grave passa a apresentar coeficientes significativos.

Entre as variáveis agrupadas, houve uma modificação na variável total_posts, deixando de ser significativa para a categoria grave, significando que postar mais não influencia na probabilidade de acesso para os usuários com níveis graves de sintomas depressivos e o coeficiente estimado para a variável tristeza, para a categoria moderado, também deixa de ser significativo.

As variáveis t1 e t4 na base posts na probabilidade de acesso à rede social para as categorias leve e grave eram superiores a categoria mínimo em comparando com a base de likes, e se tornam inferiores, permitindo concluir que os usuários com sintomas depressivos leves e graves curtem mais páginas nos primeiros e quartos trimestres do ano, entretanto realizam menos postagens considerando o mesmo período.

Afirma-se também de que as mulheres possuem mais sintomas de depressão grave quando comparadas com os homens e que o aumento do número de postagens e de curtidas está relacionado à presença de casos mais graves de sintomas depressivos.

Os resultados para os trimestres mostram que todas as pessoas utilizam mais a rede social no primeiro trimestre, mas a probabilidade de uso se eleva com o nível de sintomas depressivos. Porém no segundo e no quarto trimestres não se observa uma relação direta entre o uso da rede social e a gravidade do quadro de sintomas depressivos.

Para ilustrar, calcula-se a probabilidade esperada do evento, considerando um usuário com 25 anos, do sexo masculino, com 35 postagens e 98 curtidas no total, com 191 amigos, que está interagindo em uma madrugada do quarto trimestre, com escala de sono e apetite = 1, e com as demais variáveis = 0, assim, a probabilidade deste usuário ter comportamento depressivo leve é igual a 0.

O teste de Wald, por exemplo, mostrou que a probabilidade de um usuário da categoria grave acessar a rede, em comparação com a categoria mínimo, se reduz em 0,1% (99,9% - 100%) com a presença de mais um amigo. De modo que o acréscimo no número de amigos desmotiva os usuários com maior nível de comportamento depressivo a acessarem a rede social.

Testes adicionais de Wald mostram que cada uma das variáveis adicionadas ao modelo explica a variável label4 de modo geral sem discriminá-la em categorias, as variáveis relacionadas ao agrupamento (total_friends, total_likes e total_posts) e aos trimestres do ano também são significativas, indicando que existe uma relação entre a interação do usuário com a rede social e o período do ano com nas diferentes categorias (label4). O mesmo é válido para as variáveis idade e sexo, onde a probabilidade de acesso à rede social é diferente entre as faixas etárias e o sexo.

De maneira interdisciplinar, onde no campo da saúde obteve dados referentes ao estado emocional do usuário no tocante à sintomas depressivos, ademais, dados da rotina dos usuários da rede social Facebook, exteriorizada por meio de postagens na linha do tempo e curtidas em páginas de terceiros, foram coletados. No âmbito da mineração de dados com aporte da estatística e da aprendizagem de máquina, esses dados foram trabalhados em várias etapas até a obtenção de um Modelo de Regressão Logística Multinomial capaz de predizer com precisão as probabilidades de ocorrência de depressão em cada categoria de traços depressivos considerando o comportamento do usuário e demais variáveis.

Capítulo 5

Conclusão e Considerações finais

A coleta de dados da rede social a partir do aplicativo Vivamente e Inventário de Depressão de Beck, ambos instalados no Facebook possibilitou nesta pesquisa a extração de um conjunto de informações já adequadas ao modelo de regressão proposto.

Por meio da técnica da mineração de dados este conjunto de informações, coerentes ao objeto desta pesquisa, foram preparadas resultando na classificação da severidade depressiva dos usuários com a utilização da Escala de Depressão de Beck BDI-II.

Foi possível transformar os dados coletados em forma longitudinal por meio de um Script Mongo DB possibilitando extrair a polaridade dos sentimentos. Para que enfim, o modelo proposto fosse experimentado a partir de Test T, Coeficiente de correlação, Razão de Chance (RC) e Teste de Wald. Ressalta-se que após os testes realizados foi possível o alcance da predição de probabilidades de traços depressivos.

Isto comprova que o Modelo de Regressão Logística Multinomial, à partir do comportamento longitudinal do usuário na rede social experimentado nesta pesquisa se mostrou eficaz para a predição de probabilidades de traços depressivos.

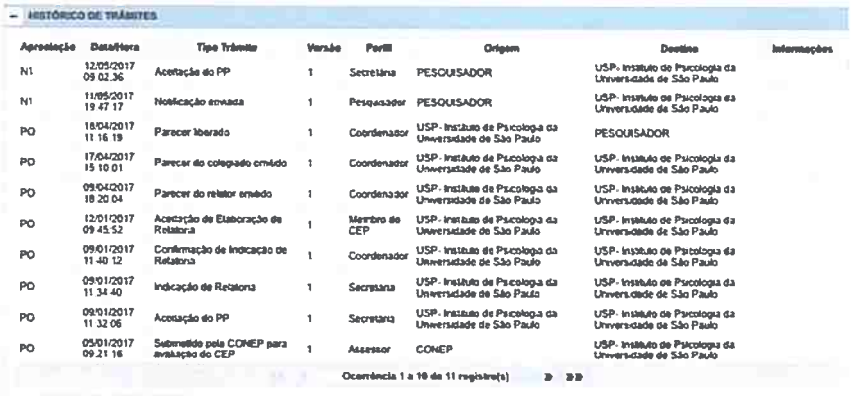
Conclui-se que as predições dos níveis de traços depressivos por meio dos atributos referentes ao comportamento do usuário na rede social representam a capacidade das ciências atuarem de maneira interdisciplinar em prol do bem estar social. Logo, o Modelo proposto pode se tornar no campo interdisciplinar mais uma ferramenta tecnológica capaz de contribuir para a saúde pública.

Apêndice A

Trâmite do projeto de pesquisa no Comitê de Ética

Detalhes da tramitação do projeto de pesquisa e seu parecer julgado pelo Comitê de Ética em Pesquisa com Seres Humanos do Instituto de Psicologia da USP (CEPH-IPUSP).

A figura A.1 exibe todo o histórico da tramitação do projeto no comitê de ética, desde a sua entrada até a aprovação no dia 12 de maio de 2017.



HISTÓRICO DE TRÂMITES

Apreciação	Data/Hora	Tipo Trâmite	Versão	Perfil	Origem	Destino	Informações
N1	12/05/2017 09:02:36	Aceitação do PP	1	Secretária	PESQUISADOR	USP- Instituto de Psicologia da Universidade de São Paulo	
N1	11/05/2017 19:47:17	Notificação enviada	1	Pesquisador	PESQUISADOR	USP- Instituto de Psicologia da Universidade de São Paulo	
PO	18/04/2017 11:16:19	Parecer liberado	1	Coordenador	USP- Instituto de Psicologia da Universidade de São Paulo	PESQUISADOR	
PO	17/04/2017 15:10:01	Parecer do colegiado emitido	1	Coordenador	USP- Instituto de Psicologia da Universidade de São Paulo	USP- Instituto de Psicologia da Universidade de São Paulo	
PO	09/04/2017 18:20:04	Parecer do relator emitido	1	Coordenador	USP- Instituto de Psicologia da Universidade de São Paulo	USP- Instituto de Psicologia da Universidade de São Paulo	
PO	12/01/2017 09:45:52	Aceitação de Elaboração de Relatoria	1	Membro do CEP	USP- Instituto de Psicologia da Universidade de São Paulo	USP- Instituto de Psicologia da Universidade de São Paulo	
PO	09/01/2017 11:40:12	Confirmação de Indicação de Relatoria	1	Coordenador	USP- Instituto de Psicologia da Universidade de São Paulo	USP- Instituto de Psicologia da Universidade de São Paulo	
PO	09/01/2017 11:34:40	Indicação de Relatoria	1	Secretaria	USP- Instituto de Psicologia da Universidade de São Paulo	USP- Instituto de Psicologia da Universidade de São Paulo	
PO	09/01/2017 11:32:06	Aceitação do PP	1	Secretaria	USP- Instituto de Psicologia da Universidade de São Paulo	USP- Instituto de Psicologia da Universidade de São Paulo	
PO	05/01/2017 09:21:16	Submetido pela CONEP para avaliação do CEP	1	Assessor	CONEP	USP- Instituto de Psicologia da Universidade de São Paulo	

Ocorrência 1 a 10 de 11 registro(s)

Figura A.1: *Histórico da tramitação do projeto de pesquisa*

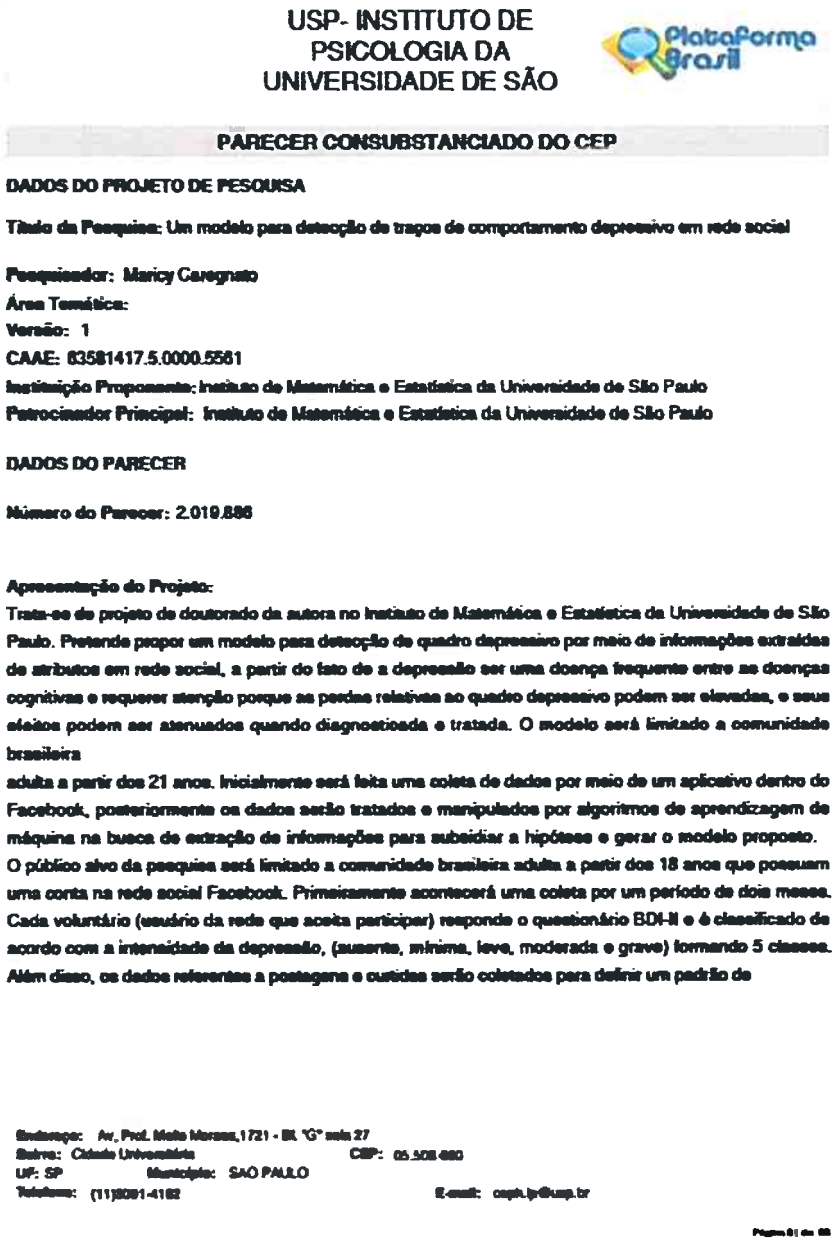
A figura A.2 mostra o estado do projeto em que no campo "situação"encontra-se aprovado.

LISTA DE PROJETOS DE PESQUISA:

Tipo	CAAE	Versão	Pesquisador Responsável	Comitê de Ética	Instituição	Origem	Última Apreciação	Situação	Ação
P	63581417.5.0000.5561	1	Mariby Canegrate	5561 - USP- Instituto de Psicologia da Universidade de São Paulo	Instituto de Matemática e Estatística da Universidade de São Paulo	PO	PO	Aprovado	

Figura A.2: *Estado de aprovado do projeto de pesquisa.*

Após a aprovação do projeto, ocorreu a liberação do Parecer favorável, conforme exposto a seguir:
As figuras A.3 à A.7 exibem o parecer completo, liberado pelo comitê de ética.



USP- INSTITUTO DE PSICOLOGIA DA UNIVERSIDADE DE SÃO

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Um modelo para detecção de traços de comportamento depressivo em rede social

Pesquisador: Maricy Caregnato
Área Temática:
Versão: 1
CAAE: 63581417.5.0000.5561
Instituição Proponente: Instituto de Matemática e Estatística da Universidade de São Paulo
Patrocinador Principal: Instituto de Matemática e Estatística da Universidade de São Paulo

DADOS DO PARECER

Número do Parecer: 2.019.886

Apresentação do Projeto:
Trata-se de projeto de doutorado da autora no Instituto de Matemática e Estatística da Universidade de São Paulo. Pretende propor um modelo para detecção de quadro depressivo por meio de informações extraídas de atributos em rede social, a partir do fato de a depressão ser uma doença frequente entre as doenças cognitivas e requerer atenção porque as perdas relativas ao quadro depressivo podem ser elevadas, e seus efeitos podem ser atenuados quando diagnosticada e tratada. O modelo será limitado a comunidade brasileira
adulta a partir dos 21 anos. Inicialmente será feita uma coleta de dados por meio de um aplicativo dentro do Facebook, posteriormente os dados serão tratados e manipulados por algoritmos de aprendizagem de máquina na busca de extração de informações para subsidiar a hipótese e gerar o modelo proposto.
O público alvo da pesquisa será limitado a comunidade brasileira adulta a partir dos 18 anos que possuem uma conta na rede social Facebook. Primeiramente acontecerá uma coleta por um período de dois meses. Cada voluntário (usuário da rede que aceita participar) responde o questionário BDI-II e é classificado de acordo com a intensidade da depressão, (ausente, mínima, leve, moderada e grave) formando 5 classes. Além disso, os dados referentes a postagens e curtidas serão coletados para definir um padrão de

Endereço: Av. Prof. Mello Moraes,1721 - Bl. "G" sala 27
Bairro: Cidade Universitária CEP: 05.508-900
UF: SP Município: SAO PAULO
Telefone: (11)3091-4182 E-mail: ceph.ip@usp.br

Página 01 de 05

Figura A.3: *Parecer completo, liberado pelo comitê de ética - parte 1*

USP- INSTITUTO DE PSICOLOGIA DA UNIVERSIDADE DE SÃO

Continuação do Parecer: 2.019.888

comportamento relacionado a sua frequência de atividades. Posteriormente à etapa da coleta de dados, ocorrerá a fase de análise, onde técnicas de mineração de dados, estatísticas e de aprendizado de máquina serão utilizadas. Os dados serão manipulados por essas técnicas onde cada classe de voluntários terá seu padrão de
comportamento e esse padrão será comparado para verificar se pode ser confirmada a hipótese de que é possível identificar correlações entre a classe de depressão e o padrão de comportamento na rede social caracterizado por postagens e curtidas.

Objetivo da Pesquisa:

Objetivo Primário:

Possibilitar a detecção de possíveis comportamentos depressivos pelas correlações entre uma classe de depressão e o seu padrão de comportamento na rede social caracterizado por postagens e curtidas.

Objetivo Secundário:

Apresentar resultados da utilização de diferentes algoritmos de aprendizagem de máquina em problemas que envolvam análise em dados de redes sociais.

Mostrar se o modelo BDI-II - Inventário de depressão de Beck será eficiente na detecção de comportamento depressivo em redes sociais.

Avaliação dos Riscos e Benefícios:

Sobre os Riscos, os pesquisadores esclarecem: "Existe a possibilidade do participante sentir-se desinformado quanto aos objetivos da pesquisa, podendo gerar um sentimento de insegurança. Para eliminar essa possibilidade um documento com informações detalhadas foi criado e disponibilizado no link: http://projetovivamente.blogspot.com.br/2016/10/sobre-o-aplicativo-e-instrucoes-de-uso.html

No que tange a forma de preenchimento do formulário, pode acontecer de o participante sentir-se despreparado com relação a utilização do aplicativo de coleta, gerando assim certo grau de ansiedade. Para resolver esse problema as instruções de uso foram detalhadamente descritas em um documento e disponibilizadas no link: http://projetovivamente.blogspot.com.br/2016/10/sobre-o-aplicativo-e-instrucoes-de-uso.html

Referente a Política de privacidade, Termos de serviço e Suporte ao usuário. Existe uma probabilidade de que o participante sinta-se desamparado, para eliminar esse sentimento cada um desses aspectos foi descrito nos links:

http://projetovivamente.blogspot.com.br/2016/10/politica-de-privacidade.html

http://projetovivamente.blogspot.com.br/2016/10/termos-de-servico.html

Endereço: Av. Prof. Mello Moraes,1721 - Bl. "G" sala 27
Bairro: Cidade Universitária CEP: 05.508-030
UF: SP Município: SAO PAULO
Telefone: (11)3091-4182 E-mail: ceph.ip@usp.br

Figura A.4: *Parecer completo, liberado pelo comitê de ética - parte 2*

USP- INSTITUTO DE PSICOLOGIA DA UNIVERSIDADE DE SÃO

Continuação do Parecer: 2.019.886

http://projetovivamente.blogspot.com.br/2016/10/informacoes-de-suporte-ao-usuario-nome.html
Respectivamente.
Em relação à coleta, armazenamento e manipulação dos dados, esse aspecto pode gerar um sentimento de insegurança referente a exposição dos dados do participante. Para suprimir esses riscos um aplicativo foi desenvolvido sob tecnologias seguras e está operando do lado do servidor; os dados serão armazenados em um banco de dados seguro e pago. E a forma de manipulação dos dados está descrita no link:
http://projetovivamente.blogspot.com.br/2016/10/sobre-o-aplicativo-e-instrucoes-de-uso.html
O participante também terá a garantia de que a manipulação de seus dados jamais permitirá sua identificação, descrita no Termo de Consentimento Livre e Esclarecido. Sob esse mesmo aspecto, existe a garantia de que o aplicativo de coleta deverá passar pela rigorosa análise de aplicativos e, quando aprovado, seguir a Política da Plataforma do Facebook: https://developers.facebook.com/policy
Uma cópia do código fonte do aplicativo encontra-se no github no link:
https://github.com/omaricy/Vivamente"
A respeito dos Benefícios:
"O participante estará contribuindo para a comunidade acadêmica de maneira geral que busca sempre técnicas mais adequadas em busca de melhores resultados, bem como o benefício a toda a comunidade que sofre com problemas de depressão.
A comunidade acadêmica, mais especificamente a comunidade que trabalha com pesquisas relativas à mineração de dados em redes sociais, será beneficiada com os resultados dos experimentos, pois é uma área onde existe uma infinidade de ferramentas e técnicas para extrair informação relevante das mais variadas bases de dados para diversos objetivos.
A comunidade de desenvolvedores de aplicativos em redes sociais, pois terão um modelo de aplicativo disponível na rede que poderá ser adaptado a objetivos semelhantes.
A comunidade acadêmica relacionada a área da saúde estará se beneficiando pois poderá manipular uma ferramenta onde poderá ter resultados relativos a um problema que é a depressão, ao qual podemos destacar:
-Do ponto de vista corporativo, identificar precocemente possíveis estados de depressão, evitando a perda de produtividade por parte do empregado, pois a pessoa que apresenta um quadro depressivo considerado grave tem sua capacidade social e produtiva comprometida em 90%; em casos moderados, 40%, e em casos com sintomas leves, 20%.

Endereço: Av. Prof. Mello Moraes,1721 - Bl. "G" sala 27
Bairro: Cidade Universitária CEP: 05.508-030
UF: SP Município: SAO PAULO
Telefone: (11)3091-4182 E-mail: cep.ip@usp.br

Página 09 de 10

Figura A.5: *Parecer completo, liberado pelo comitê de ética - parte 3*

USP- INSTITUTO DE PSICOLOGIA DA UNIVERSIDADE DE SÃO

Continuação do Parecer: 2.016.666

-Do ponto de vista econômico, diminuir o custo do tratamento da depressão, tanto para o doente quanto para a sociedade em geral. O custo relacionado com morte prematura e o custo indireto por redução na produtividade e absenteísmo no trabalho é elevado, porém quanto mais cedo a doença for diagnosticada, menos grave ela será, e consequentemente, os resultados do tratamento surtirão efeito mais cedo, diminuindo esse custo.
-Do ponto de vista psíquico, evitar perdas como a da memória, pois as pessoas que sofrem de depressão possuem alterações cerebrais quando comparadas às que não apresentam o quadro. O hipocampo, uma pequena parte do cérebro responsável pelo armazenamento da memória, é menor em pessoas que tiveram depressão.

Comentários e Considerações sobre a Pesquisa:
Pesquisa de relevância social, bem delineada, apresenta detalhadamente os riscos e providências a serem tomadas, bem como os benefícios sociais da pesquisa.

Considerações sobre os Termos de apresentação obrigatória:
Os Termos de apresentação obrigatória são apresentados adequadamente de modo a permitir a análise ética do projeto. o TCLE está adequado e a pesquisa se dará exclusivamente em ambiente virtual, pela rede social do facebook.

Recomendações:
Sem recomendações.

Conclusões ou Pendências e Lista de Inadequações:
O projeto está aprovado com a recomendação de incluir nas informações destinadas aos participantes a referência a serviços públicos de saúde mental.

Considerações Finais a critério do CEP:
Se o projeto prevê aplicação de TCLE, todas as páginas do documento deverão ser rubricadas pelo pesquisador e pelo voluntário e a última página assinada por ambos, conforme Carta Circular no 003/2011 da CONEP/CNS.
Salientamos que o pesquisador deve desenvolver a pesquisa conforme delineada no protocolo aprovado.
Eventuais modificações ou emendas ao protocolo devem ser apresentadas ao CEPH de forma clara e sucinta, identificando a parte do protocolo a ser modificada e suas justificativas. Lembramos que esta modificação necessitará de aprovação ética do CEPH antes de ser implementada. De acordo com a Res. CNS 466/12, o pesquisador deve apresentar a este CEP/SMS o relatório final do projeto desenvolvido, conforme preenchimento de Protocolo disponível na página do Comitê de Ética em

Endereço: Av. Prof. Mello Moraes,1721 - Bl. "G" sala 27
Bairro: Cidade Universitária CEP: 05.508-030
UF: SP Município: SAO PAULO
Telefone: (11)3091-4182 E-mail: ceph.ip@usp.br

Figura A.6: *Parecer completo, liberado pelo comitê de ética - parte 4*

USP- INSTITUTO DE PSICOLOGIA DA UNIVERSIDADE DE SÃO

Continuação do Parecer: 2.016.888

Pesquisa com Seres Humanos do IPUSP, do site do IPUSP. Em seguida, o protocolo preenchido deverá ser enviado ao CEPH pela Plataforma Brasil, ícone Notificação, logo que o mesmo estiver concluído.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_850505.pdf	04/01/2017 23:47:25		Aceito
Projeto Detalhado / Brochura Investigador	Projeto_de_pesquisa.pdf	04/01/2017 23:45:30	Marcy Caregnato	Aceito
Outros	Declaracao_compromisso_com_resultados_Marcy_Caregnato.pdf	04/01/2017 17:22:45	Marcy Caregnato	Aceito
Declaração de Instituição e Infraestrutura	Demonstrativo_infraestrutura.pdf	04/01/2017 17:19:04	Marcy Caregnato	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLE_Marcy_Caregnato.pdf	04/01/2017 17:11:17	Marcy Caregnato	Aceito
Folha de Rosto	Folha_de_rosto_Marcy_Caregnato.pdf	04/01/2017 17:04:01	Marcy Caregnato	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

SAO PAULO, 18 de Abril de 2017

Assinado por:
Helena Rinaldi Rosa
(Coordenador)

Endereço: Av. Prof. Mello Moraes,1721 - Bl. "G" sala 27
Bairro: Cidade Universitária CEP: 05.508-030
UF: SP Município: SAO PAULO
Telefone: (11)3091-4182 E-mail: ceph.ip@usp.br

Página 05 de 05

Figura A.7: *Parecer completo, liberado pelo comitê de ética - parte 5*

Apêndice B

Criação e configuração do app Vivamente

B.1 Configuração do Facebook Canvas

Primeiramente foi necessário dispor de uma conta na rede social Facebook, e em seguida entrar na página de *developers*[1], clicar no link *"adicionar um novo aplicativo"*. Onde foram inicialmente solicitados, o nome, o e-mail e a categoria do aplicativo (*app*). Na sequência, no painel de controle da página de configuração do *app*, onde foram automaticamente fornecidas a versão, o ID e a chave secreta do *app*, como pode ser visualizado na figura B.1.

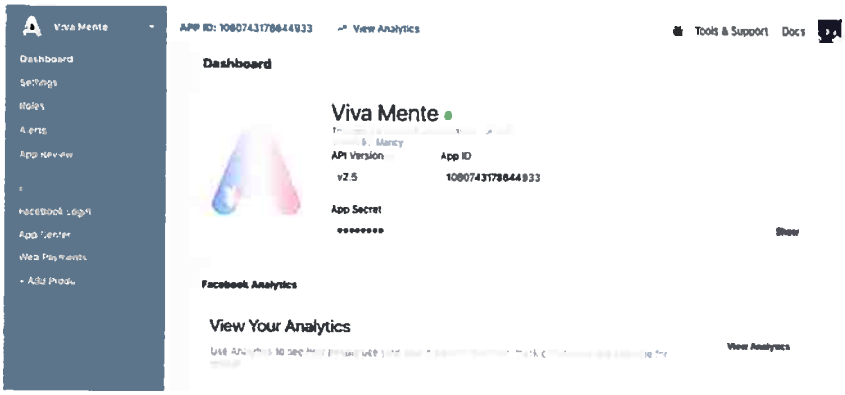


Figura B.1: *Painel de controle do Facebook canvas para o app Vivamente*

Em seguida no menu *configurações* > *básico*, foram preenchidas as informações referentes ao domínio do aplicativo[2], que é o domínio básico do servidor onde o *app* está hospedado, nesse caso o servidor *Heroku*; a URL da política de privacidade[3], que por exigência do Facebook é necessário fornecer ao usuário todas as garantias de sigilo e confidencialidade dos dados coletados; *URL* dos termos de serviço, que é uma página onde estão dispostos os termos aos quais o usuário está concordando antes de prosseguir no *app*; a *URL* do canvas seguro[4]. A página com o *link* completo que obrigatoriamente deverá ser segura, (o app deverá estar hospedado em um servidor seguro (https)), a página do canvas[5], que é o endereço da página configurado dentro do *Facebook canvas*, (utilizada para iniciar a navegação no *app* Vivamente). A figura B.2 exibe esse processo.

[1] frozen-thicket-68161.herokuapp.com
[2] http://projetovivamente.blogspot.com.br/2016/10/politica-de-privacidade.html
[3] https://frozen-thicket-68161.herokuapp.com/auth/facebook/canvas/
[4] https://apps.facebook.com/vivamente
[5] https://github.com/cmaricy/Vivamente

Figura B.2: *Configurações básicas do Facebook canvas para o app Viva Mente*

B.2 Desenvolvimento da parte Web do aplicativo Vivamente

Como dito anteriormente foi necessário escolher as ferramentas necessárias para o desenvolvimento do *app*, que após a preparação do Facebook Canvas, foi necessário desenvolver a parte Web a ser embutida no Canvas, seguindo alguns critérios de desenvolvimento.

Iniciando pela linguagem de programação, o Facebook oferece uma série de *Software Development Kits (SDKs)* oficiais, ou seja, pacotes para integração do projeto Web com o *Facebook canvas*, possibilitando trabalhar com sua *Graph API*.

A Graph API do Facebook é um serviço do tipo RESTful que retorna arquivos do tipo JSON. O processo é feito através do envio de uma solicitação HTTP para iniciar uma conexão com o Facebook, o que permite executar métodos GET para retornar dados que poderão ser posteriormente analisados.

Para o desenvolvimento do projeto web optou-se por um *SDK Javascript*. O projeto contém todo o código, com as extensões *JS, EJS* e *JSON*, necessário.

Como o Facebook exige que o *app*, para ser embutido no *canvas*, deva ser desenvolvido sob o paradigma *server side*, os *scripts JavaScript* puros não atenderam as necessidades totais por serem *client side*, sendo necessária uma tecnologia segura, onde o *NodeJS* foi selecionado.

Node.js é um interpretador de código JavaScript que funciona do lado do servidor. Juntamente com o *NodeJS* as estruturas *express, passport, angularJS* e *MongoDB* foram utilizadas.

O projeto completo com todo o código fonte está armazenado no *github*, no repositório sob o nome de Vivamente[6]. Como pode ser visualizado e resumidamente explicado após a figura B.3.

[6]http://www.ip.usp.br/

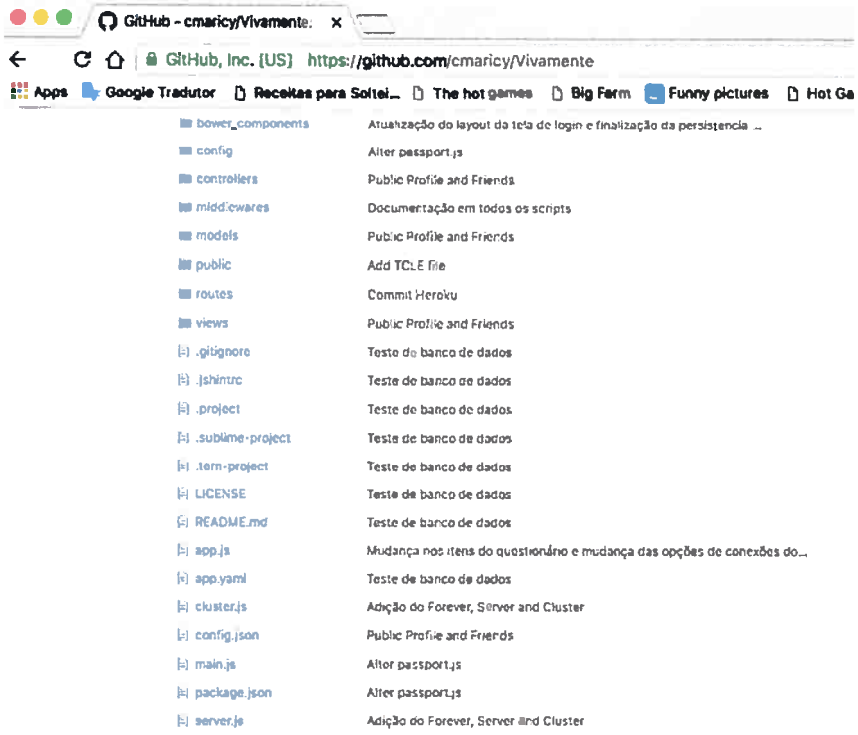


Figura B.3: *Código fonte do projeto web armazenado no Github*

O arquivo *questionário.ejs*, (*ejs* é um *framework* responsável pela renderização de HTML dinâmico com *Node.js*) no diretório *views*, define a interface gráfica da página inicial do aplicativo, ou seja, algumas informações iniciais juntamente com as 21 questões do questionário DBI-II.

O arquivo de *script*, *app.js*, gerado pelo framework *express*, é o local onde foram carregadas todas as configurações necessárias, fazendo todos os *requires*, dentre os principais destacam-se o *framework espress*, o *midleware passport* e a biblioteca *mongoose* para o banco de dados.

No diretório *config*, estão 3 dos mais importantes arquivos do aplicativo, o *auth.js*, que contém o *script* responsável por fornecer o objeto com os parâmetros de autorização do Facebook, como o *clientID* e *ClientSecret*. O *database.js que* faz a exportação da URI de conexão com *MongoDB*, configurando o nome da base de dados online (*modulus*), que está hospedando a base de dados *MongoDB*, uma pequena amostra do código desses 2 arquivos pode ser visualizado na figura B.4.

R_	label2	label4	nivel	choro	culpa	critica	fracasso	apetite	estima	suicida
1	sem_depr	minimo	8	0	0	0	0	0	0	0
2	sem_depr	minimo	8	0	0	0	0	0	0	0
3	sem_depr	minimo	8	0	0	0	0	0	0	0
4	sem_depr	minimo	8	0	0	0	0	0	0	0
5	sem_depr	minimo	8	0	0	0	0	0	0	0
6	sem_depr	minimo	8	0	0	0	0	0	0	0
7	sem_depr	minimo	8	0	0	0	0	0	0	0
8	sem_depr	minimo	8	0	0	0	0	0	0	0
9	sem_depr	minimo	8	0	0	0	0	0	0	0
10	sem_depr	minimo	8	0	0	0	0	0	0	0
11	sem_depr	minimo	8	0	0	0	0	0	0	0
12	sem_depr	minimo	8	0	0	0	0	0	0	0
13	sem_depr	minimo	8	0	0	0	0	0	0	0
14	sem_depr	minimo	8	0	0	0	0	0	0	0
15	sem_depr	minimo	8	0	0	0	0	0	0	0

Figura B.4: *Amostra do código fonte dos arquivos no diretório de configuração*

Além do *auth.js* e do *database.js*, o diretório config possui o arquivo *passport.js*, este *script* é responsável por implementar funções de autenticação da biblioteca *passport.js*, dentre eles o *FacebookStrategy*, responsável por obter a estratégia (*authentication mechanisms*) do *passport-facebook-canvas*, onde implementa os métodos necessários para toda a autenticação de um usuário no Facebook, bem como a coleta de seus dados e armazenamento no banco de dados.

Os demais arquivos de *script .js* são utilizados para outras configurações. A exemplo do *fbroutes.js*, este arquivo gerencia as rotas referentes aos encaminhamentos do Facebook feito pelo *express*. O *routes.js*, gerencia as rotas referentes aos encaminhamentos do site feito pelo *express*.

Os arquivos *JSON* fazem o mapeamento das variáveis, a exemplo do *package.JSON* que faz o gerenciamento de versão.

Mongoose é uma biblioteca *NodeJS* que fornece um mapeamento a objeto *MongoDB* com uma interface familiar dentro do *NodeJS*. Ele traduz os dados do banco de dados de objetos *JavaScript* para o *app*.

Apêndice C

Solicitações de permissões para o app no Facebook.

Histórico de solicitações de permissões que o app Vivamente foi submetido a Revisão do Facebook, conforme mostra a figura C.1.

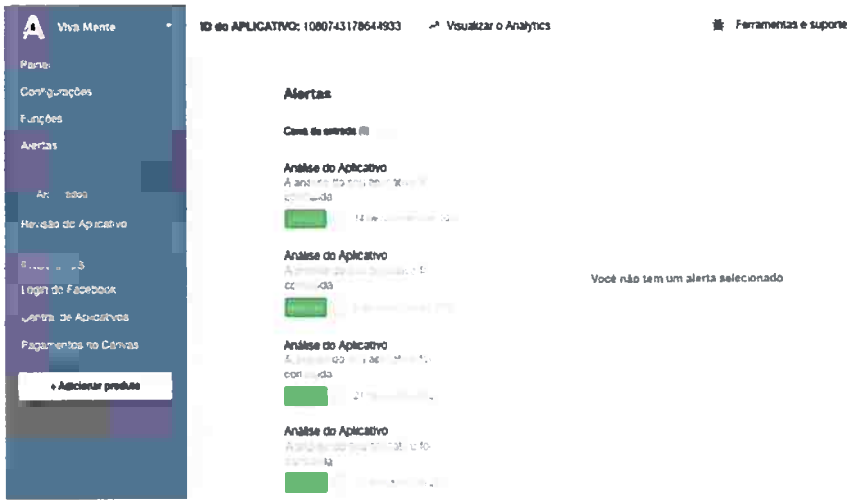


Figura C.1: *Histórico de solicitações de permissão para o app Viva Mente*

Observa-se na figura C.3, os itens aprovados pelo Facebook, tanto os aprovados por padrão, que são referentes ao perfil público, e-mail, ***user_friends*** (amigos do usuário), ***user_likes*** e ***user_posts*** (curtidas e postagens do usuário).

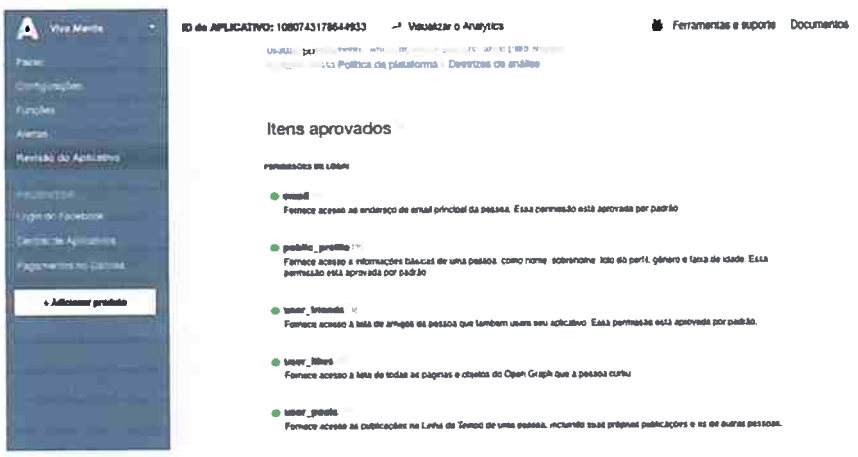


Figura C.2: *Itens aprovados pelo Facebook para o app Viva Mente*

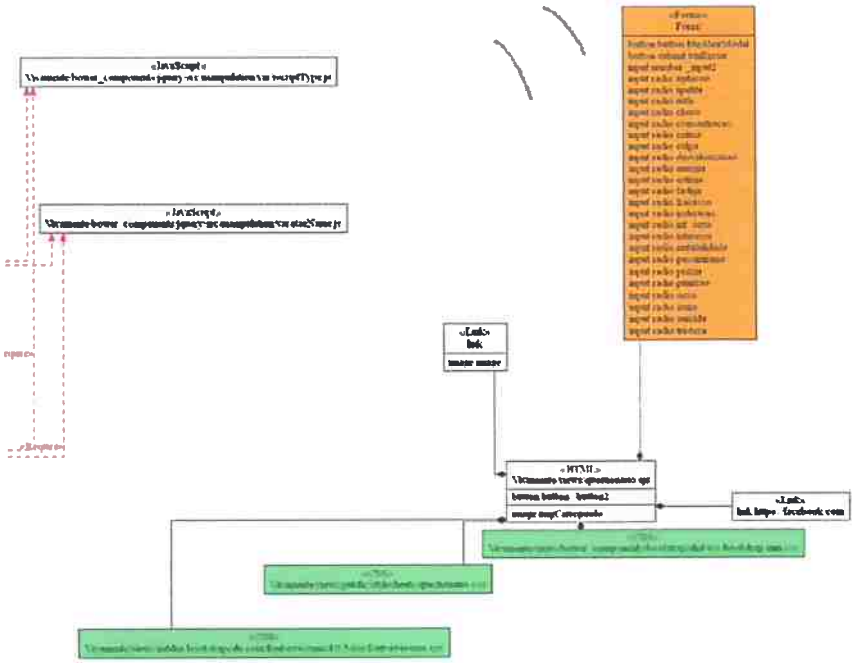


Figura C.3: *Diagrama de Classes*

Obs. O diagrama de classes completo, pelo motivo da extensão não será apresentado, porém estará disponível no github [1].

[1] https://github.com/cmaricy/Vivamente http://vivamente.herokuapp.com/

Apêndice D

Campanha de divulgação para obter Voluntários

A primeira forma de divulgação do aplicativo objetivando angariar voluntários para participar da pesquisa foi por meio do envio de um e-mail para uma lista de colegas de trabalho, amigos e familiares.

E-mail direcionado a amigos e colegas com o seguinte texto:

Olá! Você já teve depressão? Conhece alguém que teve? Essa doença é muito comum, e, a cada dia, mais pessoas sofrem desse mal. Mas o que talvez você não saiba é que pode ajudar no combate a essa doença. Para isso, basta participar desta pesquisa, que tem por objetivo identificar comportamentos depressivos pelo Facebook. É rápido, fácil e sua participação será de grande ajuda. Vamos lá!

Você só precisa ter uma conta no Facebook, estar conectado a Internet e acessar qualquer navegador.

Clique no link abaixo ou copie e cole o seguinte endereço em seu navegador:

https://apps.facebook.com/vivamente

Preencha os dados iniciais, em seguida responda as 21 questões e clique em *enviar*.

Vídeo disponível também para compartilhamento em:

Fique à vontade para consultar mais informações no blog do projeto:

Ou entre em contato por e-mail ou telefone:

11 94104 9969

Um abraço

A segunda forma de obter voluntários foi por meio de publicações e compartilhamentos nas linhas do tempo de algumas redes sociais:

No Facebook juntamente com o link, conforme mostra a figura D.1.

Figura D.1: *Exemplo de um compartilhamento na linha do tempo de um usuário no Facebook*

A figura D.2 mostra a divulgação do aplicativo no Twitter.

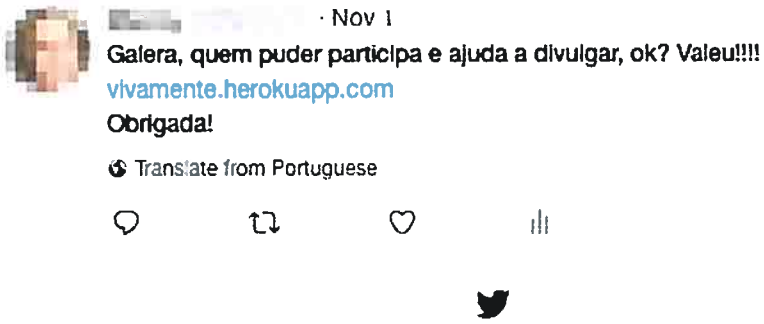


Figura D.2: *Exemplo de uma publicação na linha do tempo no Twitter*

A figura D.3 exibe um exemplo de uma publicação na linha do tempo no Linkedin

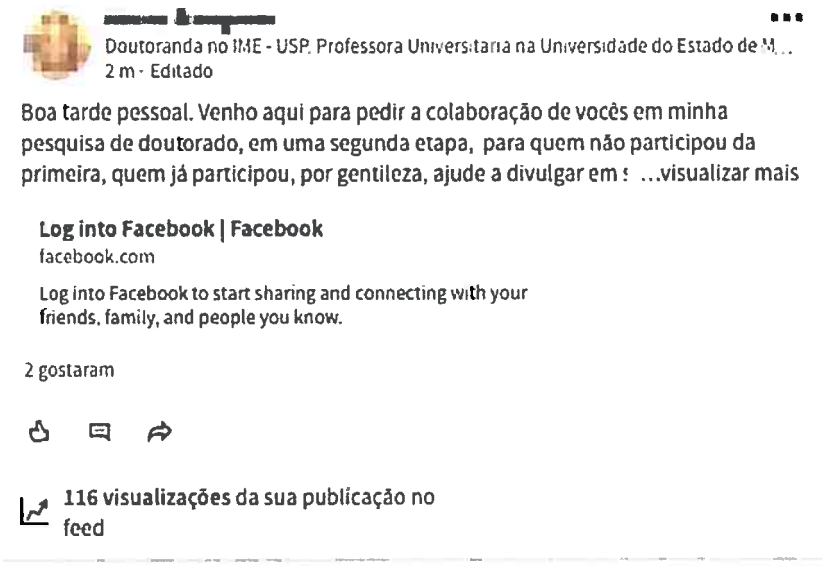


Figura D.3: *Exemplo de uma publicação na linha do tempo no Linkedin*

A terceira maneira de divulgar o aplicativo foi por meio de uma página criada no Facebook (Fan page), com informações sobre o aplicativo, tendo sido promovida para gerar maior visibilidade do *app* Vivamente, porém o alcance foi baixo, apenas 246 pessoas, conforme mostra a figura D.4.

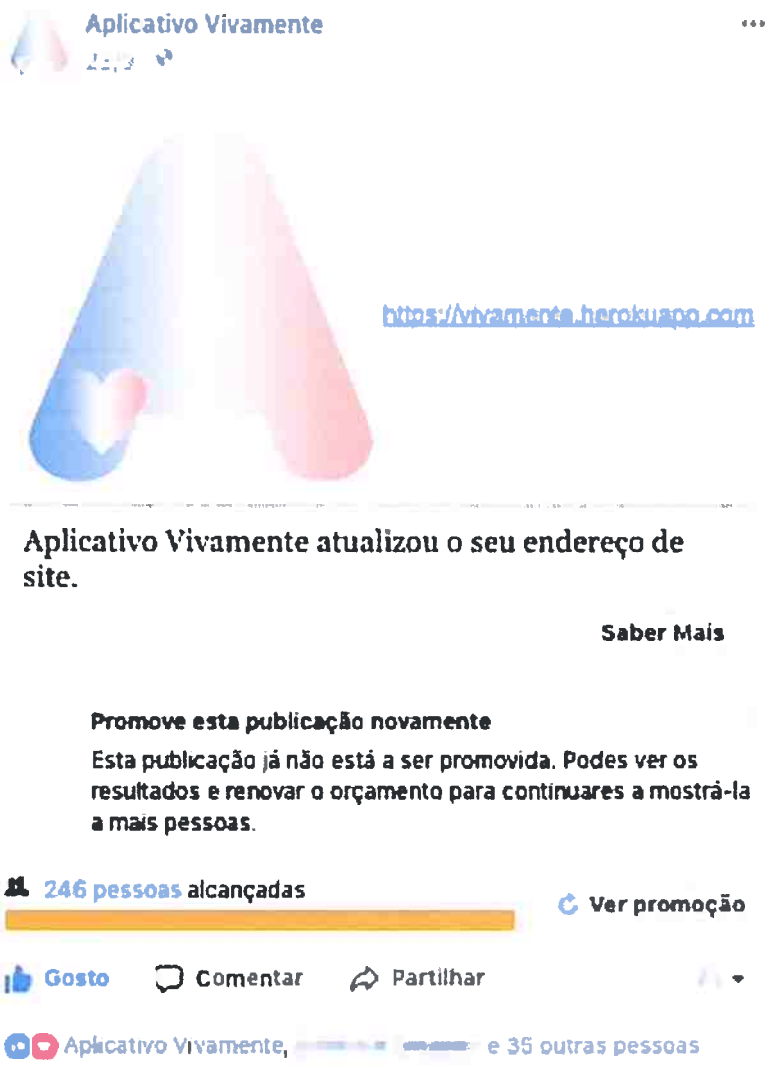


Figura D.4: *Exemplo do anúncio da página Vivamente*

A URL do vídeo pode ser visualizada em:
https://www.facebook.com/cmaricy/

A quarta forma de promoção foi uma campanha de promoção via Facebook do próprio *app* Vivamente, com o objetivo de conseguir maior número de usuários para instalar o *app*, pois nessa forma de divulgação havia a possibilidade de escolher a opção de efetuar o pagamento por instalação, onde o alcance foi bem mais alto, chegando a 17.934 pessoas, conforme pode ser visualizado na figura D.5.

Figura D.5: *Exemplo do anúncio promovendo o app Vivamente*

A figura mostra um resumo da campanha onde pode-se visualizar o alcance e também a quantidade de instalações, que foram 114, e os valore pagos conforme mostra a figura D.6.

Figura D.6: *Resultados da campanha da página Vivamente e do app Viva Mente*

A quinta forma de divulgação foi por intermédio de um vídeo criado no youtube, onde o alcance foi bem menor, 2.369 pessoas visualizaram, essa campanha também foi paga, porém teve apenas 5 instalações durante o período de vigência, a figura D.7 mostra

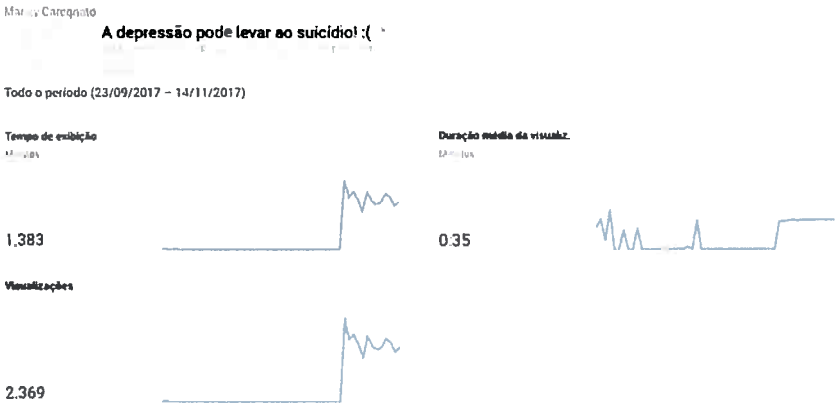


Figura D.7: *Resultados da campanha do vídeo no Youtube*

A URL do vídeo pode ser visualizada em:
http://https://youtu.be/S16_GpKqU-s/

Apêndice E

Fluxo de utilização do aplicativo Vivamente

O objetivo dessa descrição, é mostrar a nível de usuário como funciona o processo total de utilização do aplicativo de coleta Vivamente.

Primeiramente o usuário deve digitar o seguinte endereço no seu navegador, para entrar no *app* Vivamente:

https://apps.facebook.com/vivamente/

Se o usuário NÃO estiver logado no Facebook, aparecerá uma tela conforme a figura E.1, para que entre com seu usuário e senha, sendo automaticamente direcionado ao aplicativo Viva Mente.

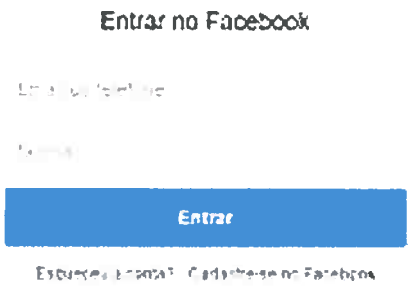


Figura E.1: *Tela inicial do Facebook*

Se o usuário estiver logado no Facebook, aparecerá uma tela conforme a figura E.2. O usuário deverá autorizar, ou seja, aceitar a tela de permissão, para que o Viva Mente acesse seus dados e após clicar no botão "*Continuar como «SeuNome»*".

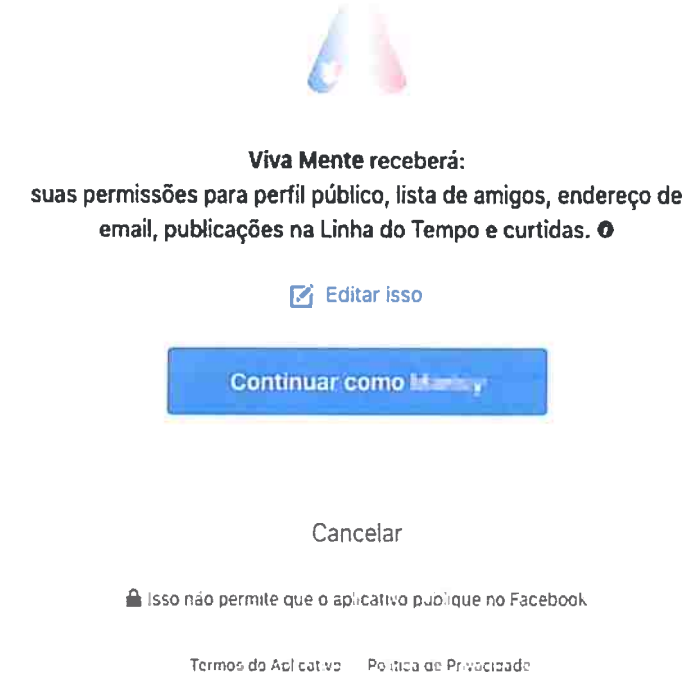


Figura E.2: *Tela de permissão ao Viva Mente*

Após conceder às permissões o usuário será direcionado à tela inicial do Viva Mente, onde deverá preencher alguns dados iniciais e em seguida responder as 21 questões do questionário BDI-II. Conforme mostram as figuras E.3 e E.4.

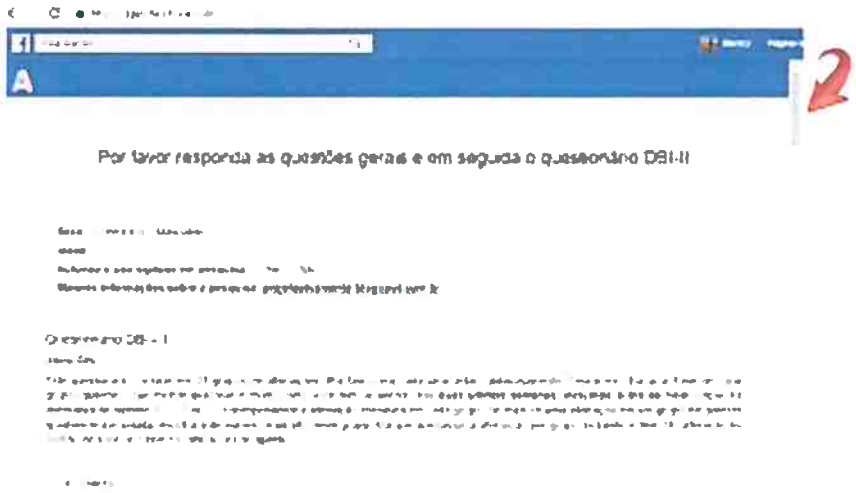


Figura E.3: *Tela inicial do questionário*

Figura E.4: *Tela final do questionário*

Após o usuário responder às questões, esses dados foram armazenados juntamente com seus dados referentes a todo o histórico de curtidas e postagens no Facebook, em um bancos de dados MongoDB.

Apêndice F

Termo De Consentimento Livre e Esclarecido - TCLE

Título da pesquisa: **Um modelo para detecção de traços de comportamento depressivo em rede social.**
Pesquisadora principal: Maricy Caregnato
Pesquisador assistente: Prof. Dr. Flávio Soares Corrêa da Silva

1. **Natureza da pesquisa:** o(a) Sr.(a.) está sendo convidado(a) a participar desta pesquisa que tem como finalidade detectar possíveis comportamentos depressivos pela correlação entre uma classe de depressão e o seu padrão de comportamento na rede social Facebook caracterizado por postagens e curtidas.

Durante a pesquisa o(a) Sr.(a.) deverá digitar o endereço https://apps.facebook.com/vivamente/ em um computador pessoal de sua preferência e em um local que lhe seja mais conveniente, com acesso a internet e possuindo uma conta na rede social Facebook. Se o(a) Sr.(a.) não estiver logado no Facebook, será solicitado para que entre com seu *login* e *senha*, em seguida aparecerá uma tela solicitando permissão para coletar seus dados de postagens e curtidas, clique no botão *continuar como <<seu nome>>*. Após a página ser carregada o(a) Sr.(a.) estará no aplicativo *vivamente* onde visualizará um questionário (BDI-II) com 21 (vinte e uma) questões de múltipla escolha, que avaliará seu grau de depressão e armazenará esses dados. Leia as instruções com atenção e responda com a máxima sinceridade possível. Na página há um link com mais informações e contato dos pesquisadores.

1. **Participantes da pesquisa:** Qualquer pessoa com idade acima de 18 anos que possua uma conta na rede social Facebook.

2. **Envolvimento na pesquisa:** ao participar deste estudo o(a) Sr.(a.) permitirá que os pesquisadores responsáveis utilizem os dados coletados durante o experimento para fins acadêmicos, porém sempre preservando a sua identidade, que nunca será revelada publicamente.

   O(a) Sr.(a.) tem liberdade de se recusar a participar e ainda se recusar a continuar participando em qualquer fase da pesquisa, sem necessidade de qualquer explicação, o que não trará prejuízo algum para o(a) Sr.(a.). Sempre que quiser poderá pedir mais informações sobre a pesquisa entrando em contato com os pesquisadores do projeto.

3. **Riscos e desconforto:** a participação nesta pesquisa não traz complicações legais. Os procedimentos adotados nesta pesquisa obedecem aos Critérios da Ética em Pesquisa com Seres Humanos conforme Resolução no. 466/2012 do Conselho Nacional de Saúde. Nenhum dos procedimentos usados oferece riscos à sua saúde ou dignidade.

4. **Garantia de Indenização**: Apesar desta pesquisa não trazer complicações legais e não oferecer riscos à saúde ou à dignidade, o(a) Sr.(a.) tem a garantia de que qualquer eventual dano será devidamente indenizado pela instituição dos pesquisadores responsáveis, nas formas da lei brasileira.

5. **Confidencialidade:** Todas as informações coletadas neste estudo são estritamente confidenciais. Somente os pesquisadores responsáveis terão conhecimento dos dados do(a) Sr.(a.).

6. **Benefícios**: Ao participar desta pesquisa, o(a) Sr.(a.) não terá nenhum benefício direto. Entretanto, esperamos que este estudo resulte em informações relevantes para a identificação de traços de comportamento depressivo em redes sociais pois identificando a depressão em seu estágio inicial pode-se evitar danos maiores causada por essa doença.

7. **Pagamento:** O(a) Sr.(a.) não terá nenhum tipo de despesa para participar desta pesquisa, bem como nada será pago por sua participação.

8. **Formas de Acompanhamento**: A sua participação se resume em responder o questionário e permitir a coleta de informações referentes à suas postagens e curtidas na rede social Facebook. Os resultados desta pesquisa serão publicados após sua conclusão pela Universidade dos pesquisadores, muito embora os dados utilizados sejam confidenciais e controlados. Em caso de cancelamento do projeto, por qualquer que seja o motivo, todos os dados coletados serão imediatamente descartados de forma completa e irrecuperável. Para a sua garantia de que o que foi aqui estabelecido será devidamente cumprido, o(a) Sr.(a.) está tendo acesso a uma cópia desse documento, assinado e datado pelo pesquisador responsável ou por seu assistente autorizado – ambos devidamente identificados no final deste documento. Além disso, sempre que quiser, o(a) Sr.(a.) poderá pedir mais informações sobre a pesquisa, entrando em contato por e-mail ou por telefone com qualquer um dos pesquisadores do projeto. Caso prefira ou necessite, entre em contato com o Comitê de Ética em Pesquisa com Seres Humanos do Instituto de Psicologia da USP (CEPH-IPUSP), por meio do endereço Av. Professor Mello Moraes, 1.721, Bloco G, segundo andar, sala 27, CEP: 05508-030 Cidade Universitária - São Paulo - SP, ou por meio do e-mail ceph.ip@usp.br, ou por meio do telefone (11) 3091-4182.

Após estes esclarecimentos, solicitamos o seu consentimento de forma livre para participar desta pesquisa.
Consentimento livre e esclarecido
Tendo em vista os itens acima apresentados, eu, de forma livre e esclarecida, manifesto meu consentimento em participar da pesquisa. Declaro que recebi cópia deste termo de consentimento, e autorizo a realização da pesquisa e a divulgação dos dados obtidos neste estudo.
Assinatura do participante da pesquisa:
Assinatura do pesquisador responsável:
Pesquisadora principal: Maricy Caregnato(cmaricy@ime.usp.br)
Pesquisador assistente: Prof. Dr. Flávio Soares Corrêa da Silva(fcs@ime.usp.br)
Telefones para contato: (11) 3091-6134 e (11) 3091-6135

Apêndice G

Questionário BDI-II

Instruções:
Este questionário consiste em 21 grupos de afirmações. Por favor, leia cada uma delas cuidadosamente. Depois escolha uma frase de cada grupo (questão), que melhor descreve o modo como você tem se sentido nas **duas últimas semanas, incluindo o dia de hoje**. Clique na alternativa de número (0, 1, 2 ou 3), correspondente a afirmação escolhida em cada grupo. Se mais de uma afirmação em um grupo lhe parecer igualmente apropriada, escolha a de número mais alto neste grupo. Marque apenas uma afirmação por grupo, incluindo o item 16 (alteração no padrão de sono) e o item 18 (alterações de apetite).

1. Tristeza

   0 Não me sinto triste.

   1 Eu me sinto triste grande parte do tempo.

   2 Estou triste o tempo todo.

   3 Estou tão triste ou tão infeliz que não consigo suportar.

2. Pessimismo

   0 Não estou desanimado(a) a respeito do meu futuro.

   1 Eu me sinto mais desanimado(a) a respeito do meu futuro do que de costume.

   2 Não espero que as coisas dêem certo pra mim.

   3 Sinto que não há esperança quanto ao meu futuro. Acho que só vai piorar.

3. Fracasso passado

   0 Não me sinto um(a) fracassado(a).

   1 Tenho fracassado mais do que deveria.

   2 Quando penso no passado vejo muitos fracassos.

   3 Sinto que como pessoa sou um fracasso total.

4. Perda de prazer

   0 Continuo sentindo o mesmo prazer que sentia com as coisas de que eu gosto.

   1 Não sinto tanto prazer com as coisas como costumava sentir.

   2 Tenho muito pouco prazer nas coisas que eu costumava gostar.

   3 Não tenho mais nenhum prazer nas coisas que eu costumava gostar.

5. Sentimento de culpa

   0 Não me sinto particularmente culpado(a).

   1 Eu me sinto culpado(a) a respeito de várias coisas que fiz e/ou que deveria ter feito.

   2 Eu me sinto culpado(a) a maior parte do tempo.

   3 Eu me sinto culpado(a) o tempo todo.

6. Sentimento de punição

   0 Não sinto que estou sendo punido(a).

   1 Sinto que posso ser punido(a).

   2 Eu acho que serei punido(a).

   3 Sinto que estou sendo punido(a).

7. Auto-estima

   0 Eu me sinto como sempre me senti em releção a mim mesmo(a).

   1 Perdi a confiança em mim mesmo(a).

   2 Estou desapontado(a) comigo mesmo(a).

   3 Não gosto de mim.

8. Autocrítica

   0 Não me critico nem me culpo mais do que o habitual.

   1 Estou sendo mais crítico(a) comigo mesmo(a) do que contumava ser.

   2 Eu me critico por todos os meus erros.

   3 Eu me culpo por tudo de ruim que acontece.

9. Pensamentos ou desejos suicidas

   0 Não tenho nenhum pensamento de me matar.

   1 Tenho pensamentos de me matar, mas não levaria isso adiante.

   2 Gostaria de me matar.

   3 Eu me mataria de tivesse oportunidade.

10. Choro

    0 Não choro mais do que chorava antes.

    1 Choro mais agora do que costumava chorar.

    2 Choro por qualquer coisinha.

    3 Sinto vontade de chorar, mas não consigo.

11. Agitação

    0 Não me sinto mais inquieto(a) ou agitado(a) do que me sentia antes.

    1 Eu me sinto mais inquieto(a) ou agitado(a) do que me sentia antes.

    2 Eu me sinto tão inquieto(a) ou agitado(a) que é difícil ficar parado(a).

    3 Estou tão inquieto(a) ou agitado(a) que tenho que estar sempre me mexendo ou fazendo alguma coisa.

12. Perda de interesse

    0 Não perdi o interesse por outras pessoas ou por minhas atividades.

    1 Estou menos interessado pelas outras pessoas ou coisas do que costumava estar.

    2 Perdi quase todo o interesse por outras pessoas ou coisas .

    3 É difícil me interessar por alguma coisa.

13. Indecisão

    0 Tomo minhas decisões tão bem quanto antes.

    1 Acho mais difícil tomar decisões agora do que antes.

    2 Tenho muito mais dificuldade em tomar decisões agora do que antes.

    3 Tenho dificuldade para tomar qualquer decisão.

14. Desvalorização

    0 Não me sinto sem valor.

    1 Não me considero hoje tão útil ou não me valorizo como antes.

    2 Eu me sinto com menos valor quando me comparo com outras pessoas.

    3 Eu me sinto completamente sem valor.

15. Falta de energia

    0 Tenho tanta energia hoje como sempre tive.

    1 Tenho menos energia do que costumava ter.

    2 Não tenho energia suficiente para fazer muita coisa.

    3 Não tenho energia suficiente para nada.

16. Alteração no padrão de sono

    0 Não percebi nenhuma mudança no meu sono.

    1a Durmo um pouco mais do que o habitual.

    1b Durmo um pouco menos do que o habitual.

    2a Durmo muito mais do que o habitual.

    2b Durmo muito menos do que o habitual.

    3a Durmo a maior parte do dia.

    3b Acordo 1 ou 2 horas mais cedo e não consigo voltar a dormir.

17. Irritabilidade

    0 Não estou mais irritado(a) do que o habitual.

    1 Estou mais irritado(a) do que o habitual.

    2 Estou muito mais irritado(a) do que o habitual.

    3 Fico irritado(a) o tempo todo.

18. Alteração de apetite

    0 Não percebi nenhuma mudança no meu apetite.

    1a Meu apetite está um pouco menor do que o habitual.

    1b Meu apetite está um pouco maior do que o habitual.

    2a Meu apetite está muito menor do que antes.

    2b Meu apetite está muito maior do que antes.

    3a Não tenho nenhum apetite.

    3b Quero comer o tempo todo.

19. Dificuldade de concentração

    0 Posso me concentrar tão bem quanto antes.

    1 Não posso me concentrar tão bem quanto habitualmente.

    2 É muito difícil manter a concentração em alguma coisa por muito tempo.

    3 Eu acho que não consigo me concentrar em nada.

20. Cansaço ou fadiga

    0 Não estou mais cansado(a) ou fadigado(a) do que o habitual.

    1 Fico cansado(a) ou fadigado(a) mais facilmente do que o habitual.

    2 Eu me sinto muito cansado(a) ou fadigado(a) para fazer muitas das coisas que costumava fazer.

    3 Eu me sinto muito cansado(a) ou fadigado(a) para fazer a maioria das coisas que costumava fazer.

21. Perda de interesse por sexo

0 Não notei qualquer mudança recente no meu interesse por sexo.
1 Estou menos interessado(a) em sexo do que costumava estar.
2 Estou muito menos interessado(a) em sexo agora.
3 Perdi completamente o interesse por sexo.
Muito obrigada pela sua cooperação!

Apêndice H

Atributos das Bases de dados likes e posts

H.0.1 Atributos da base de dados likes

_id	**identificador único**
data	**data completa com dia, mes e ano**
qtd_ativ	quantidade de vezes que executou uma atividade
label2	com depressão e sem depressão (0 ou 1)
label4	nível de depressão minimo, leve, moderado ou grave (1, 2, 3 ou 4)
ano	ano que aconteceu a ação
dia	dia que aconteceu a ação (varia de 1 a 30/31, varia o mês)
dia_ano	dia_ano que aconteceu a ação (varia de 1 a 364)
dia_semana	dia_semana que aconteceu a ação (varia de 1 a 7)
hora	hora que aconteceu a ação (varia de 0 a 24)
manha	período da manhã (0 ou 1)
tarde	período da tarde (0 ou 1)
mês	mês que aconteceu a ação (varia de 1 a 12)
minutos	minuto que aconteceu a ação (varia de 0 a 60)
segundos	minuto que aconteceu a ação (varia de 0 a 60)
qtd_likes	likes no período (ex, se for mês, será a quantidade naquele mês)
agitação	sintoma depressivo (de 0 a 3)
apetite	sintoma depressivo (de 0 a 3)
choro	sintoma depressivo (de 0 a 3)
concentração	sintoma depressivo (de 0 a 3)
critica	"
culpa	"
desvalorizacao	"
energia	"
estima	"
fadiga	"
fracasso	"
indecisao	"
int_sexo	"
interesse	"
irritabilidade	"
pessimismo	"

prazer	"
punicao	"
sono	"
suicida	"
tristeza	"
total_friends	quantidade total de amigos (desde o dia que criou o facebook até a data da coleta)
total_posts	quantidade total de posts (desde o dia que criou o facebook até a data da coleta)
total_likes	quantidade total de likes (desde o dia que criou o facebook até a data da coleta)
idade	idade do usuário
sexo	sexo do usuário (0 para mulher e 1 para homem)
data	data completa ex. 25/01/2013
noite	período da noite (0 ou 1)
segunda	se for segunda-feira (0 ou 1)
terca	"
quarta	"
quinta	"
sexta	"
sabado	"
domingo	"
madrugada	"
ano2011	"
ano2012	"
ano2013	"
ano2014	"
ano2015	"
ano2016	"
ano2017	"
finalsemana	"
janeiro	"
fevereiro	"
marco	"
abril	"
maio	"
junho	"
julho	"
agosto	"
setembro	"
outubro	"
novembro	"
dezembro	"
t1	primeiro trimestre do ano (0 ou 1)
t2	segundo trimestre do ano

t3	terceiro trimestre do ano
t4	quarto trimestre do ano

Os atributos que estão com "(aspas duplas) na descrição ou em branco é porque segue o padrão da anterior, sempre sendo o 0 (zero) como a não ocorrência do evento e o 1 (um) para a ocorrência. (sem/com, não/sim, falso/verdadeiro, negativo/positivo, etc)

H.0.2 Atributos da base de dados posts

_id	identificador único
data	data completa com dia, mes e ano
qtd_ativ	quantidade de vezes que executou uma atividade
label2	com depressão e sem depressão (0 ou 1)
label4	nível de depressão minimo, leve, moderado ou grave (1,2,3 ou 4)
ano	ano que aconteceu a ação
dia	dia que aconteceu a ação (varia de 1 a 30/31, varia o mês)
dia_ano	dia_ano que aconteceu a ação (varia de 1 a 364)
dia_semana	dia_semana que aconteceu a ação (varia de 1 a 7)
hora	hora que aconteceu a ação (varia de 0 a 24)
manha	período da manhã (0 ou 1)
tarde	período da tarde (0 ou 1)
mês	mês que aconteceu a ação (varia de 1 a 12)
minutos	minuto que aconteceu a ação (varia de 0 a 60)
segundos	minuto que aconteceu a ação (varia de 0 a 60)
qtd_message	quantidade de mensagens naquele período (ex, por mês, será a quantidade naquele mês)
qtd_story	quantidade de story (ações do usuário, como fotos no mural, troca de foto de perfil, etc) sentimento, etc) naquele período
qtd_sent_neg	quantidade de sentimento negativo naquele período (referente as postagens) (0 ou 1)
qtd_sent_pos	quantidade de sentimento positivo naquele período
agitação	sintoma depressivo (de 0 a 3)
apetite	sintoma depressivo (de 0 a 3)
choro	sintoma depressivo (de 0 a 3)
concentração	sintoma depressivo (de 0 a 3)
critica	"
culpa	"
desvalorizacao	"
energia	"
estima	"
fadiga	"
fracasso	"
indecisao	"
int_sexo	"
interesse	"
irritabilidade	"
pessimismo	"
prazer	"

punicao	"
sono	"
suicida	"
tristeza	"
total_friends	quantidade total de amigos (desde o dia que criou o facebook até a data da coleta)
total_posts	quantidade total de posts (desde o dia que criou o facebook até a data da coleta)
total_likes	quantidade total de likes (desde o dia que criou o facebook até a data da coleta)
idade	idade do usuário
sexo	sexo do usuário (0 para mulher e 1 para homem)
data	data completa ex. 25/01/2013
noite	período da noite (0 ou 1)
segunda	se for segunda-feira (0 ou 1)
terca	"
quarta	"
quinta	"
sexta	"
sabado	"
domingo	"
madrugada	"
ano2011	"
ano2012	"
ano2013	"
ano2014	"
ano2015	"
ano2016	"
ano2017	"
finalsemana	"
janeiro	"
fevereiro	"
marco	"
abril	"
maio	"
junho	"
julho	"
agosto	"
setembro	"
outubro	"
novembro	"
dezembro	"
t1	primeiro trimestre do ano (0 ou 1)
t2	segundo trimestre do ano
t3	terceiro trimestre do ano

t4	quarto trimestre do ano

Apêndice I

Estimações longitudinais para dados em painel

O quadro I apresenta, de forma consolidada, as principais estimações dos modelos longitudinais de regressão para dados em painel.

Modelo	Painel	Estimação	Descrição
Linear	Curto	GEE	Estimação POLS erros-padrão robustos com agrupamento por individuo
		Efeitos Fixos	Estimação por efeitos fixos
		Efeitos Fixos	Estimação por efeitos fixos com erros-padrão robustos com agrupamento por indivíduo
		Efeitos Aleatórios	Estimação por efeitos aleatórios
		Efeitos Aleatórios	Estimação por efeitos aleatórios com erros-padrão robusto com agrupamento por indivíduo
	Longo	GEE	Estimação POLS com efeitos auto regressivos de primeira ordem AR(??)
		GEE	Estimação POLS com efeitos autorregressivos de primeira ordem AR (??)de p-éstima ordem AR (??)
		GEE	Estimação GLS c om efeitos autorregressivos de primeira ordem AR (??) e termos de erro heterocedástico
		Efeitos Fixos	Estimação por efeitos fixos com termos de erro AR (??)
		Efeitos Aleatórios	Estimação por efeitos aleatórios com termos de erro AR (??)
Não Linear	Logístico	GEE	Estimação Pooled com erros-padrão robustos com agrupamento por indivíduo
		GEE	Estimação PA com erros padrão robustos
		Efeitos Fixos	Estimação por efeitos fixos
		Efeitos Aleatórios	Estimação por efeitos aleatórios
	Poisson	GEE	Estimação Pooled com erros-padrão robustos com agrupamento por indivíduos
		GEE	Estimação PA com erros-padrão robustos
		Efeitos aleatórios	Estimação por efeitos aleatórios
	Binomial Negativo	GEE	Estimação Pooled com erros-padrão robustos com agrupamento por indivíduo
		GEE	Estimação PA com erros-padrão robustos
		Efeitos Aleatórios	Estimação por efeitos aleatórios

Referências Bibliográficas

Abella *et al.*(2017) Joan Domènech Abella, Elvira Lara1, Maria Rubio-Valera, Beatriz Olaya, Maria Victoria Moneta, Laura Alejandra, Rico-Uribe, Jose Luis Ayuso Mateos, Jordi Mundó e Josep Maria Haro. Poisson Mixture Regression Models for Heart Disease Prediction . *Psychiatry Epidemiology*. doi: 10.1007/s00127-017-1339-3. Citado na pág. 18, 68

Ahrari *et al.*(2013) Farideh Ahrari, Seyyed Hamid Salehi, Mohammad Javad Fatemi, Madjid Soltani, Shahrzad Taghavi e Roghayeh Samimi. Severity of symptoms of depression among burned patients one week after injury, using Beck Depression Inventory-II (BDI-II). *Burns*, 39(2):285–290. ISSN 03054179. doi: 10.1016/j.burns.2012.07.012. URL http://dx.doi.org/10.1016/j.burns.2012.07.012. Citado na pág. 11

Allison(2014) Paul D. Allison. *Logistic Regression Using SAS: Theory and Application*. SAS Institute, 2nd edition edição. Citado na pág. 29, 59, 84, 90

Amemiya(1981) T. Amemiya. Qualitative Response Model: A Survey . *Journal of Economic literature*. doi: 10.1002/jps.3030411125. Citado na pág. 29

Baldwin e Birtwistle(2002) D.S. Baldwin e J. Birtwistle. *An Atlas of Depression*. 2d ed. Philadelphia: Saunders. ISBN 978853625039-7. Citado na pág. 1, 8

Baltagi(2005) Badi H. Baltagi. *Econometric Analysis of Panel Data*. John Wiley and Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England. Citado na pág. 37

Baltagi(2011) Badi H. Baltagi. *Econometric*. Springer Science & Business Media, 2011. ISBN 3642200591, 9783642200595. Citado na pág. 59

Beck e Alford(2009) Aaron T. Beck e Brad A. Alford. *Depression: Causes and Treatment*. University of Pennsylvania Press. Citado na pág. 5, 6, 7, 8

Beck e Alford(2011) Aaron T. Beck e Brad A. Alford. *Depressão: Causas e Tratamento*. Artmed. Citado na pág. 6, 7, 8

Beck e Rush(1979) Aaron T Beck e A J Rush. *Cognitive therapy of depression*. Guilford Press. Citado na pág. 6, 7, 10

Beck *et al.*(1996) Aaron T Beck, RA Steer e GK Brown. *Manual for the Beck Depression Inventory-II*. Psychological Corporation. ISBN 978853625039-7. Citado na pág. 9, 10, 11, 58

Beers(2010) Clifford W. Beers. *A Mind That Found Itself Paperback*. readclassic. ISBN 978853625039-7. Citado na pág. 5

Bertens *et al.*(2016) Loes C.M. Bertens, Karel G.M. Moons, Frans H. Rutten, Yvonne van Mourik, Arno W. Hoes e Johannes B. Reitsma. A nomogram was developed to enhance the use of multinomial logistic regression modeling in diagnostic research. *Journal of Clinical Epidemiology*, 71:51–57. ISSN 18785921. doi: 10.1016/j.jclinepi.2015.10.016. URL http://dx.doi.org/10.1016/j.jclinepi.2015.10.016. Citado na pág. 37

Blachioa *et al.*(2016) Agata Blachioa, Aneta Przepiorkaa e Igor Panticb. Association between Facebook addiction, self-esteem and life satisfaction: A cross-sectional study. página 701–705. doi: 10.1177/107319110100800409. Citado na pág. 17

Bodroža e Jovanović(2016) Bojana Bodroža e Tamara Jovanović. Validation of the new scale for measuring behaviors of Facebook users: Psycho-Social Aspects of Facebook Use (PSAFU). *Computers in Human Behavior*, 54:425–435. ISSN 07475632. doi: 10.1016/j.chb.2015.07.032. Citado na pág. 17

Bojmela *et al.*(2016) Liad Bareket Bojmela, Simone Morana e Golan Shaharb. Strategic self-presentation on Facebook: Personal motives and audience response to online behavior. doi: 10.1177/107319110100800409. Citado na pág. 16

Brown *et al.*(1995) G P Brown, C L Hammen, M G Craske e T D Wickens. Dimensions of dysfunctional attitudes as vulnerabilities to depressive symptoms. Citado na pág. 9

Caetano(1993) Dorgival Caetano. *Classificação de Transtornos Mentais e de Comportamento da Cid-10 - Descrições Clínicas e Diretrizes diagnósticas*. Editora Artmed. ISBN 8582712073. Citado na pág. 5

Calil e Pires(1998) H. M. Calil e M. L. N. Pires. Aspectos gerais das escalas de avaliação de depressão. Citado na pág. 8

Callan *et al.*(2017) Judith A. Callan, Jesse Wright, Greg J. Siegle, Robert H. Howland e Britney B. Kepler. Use of Computer and Mobile Technologies in the Treatment of Depression. *Archives of Psychiatric Nursing*, 31(3):311–318. ISSN 08839417. doi: 10.1016/j.apnu.2016.10.002. URL http://dx.doi.org/10.1016/j.apnu.2016.10.002. Citado na pág. 15

Cameron e Trivedi(2009) Colin A Cameron e Pravin K Trivedi. *Regression Models for Categorical, Count, and Related Variables An Applied Approach*. Stata corp. Citado na pág. 34, 36, 37

Carvalho e Pianowski(2017) Lucas de Francisco Carvalho e Giselle Pianowski. Pathological personality traits assessment using Facebook: Systematic review and meta-analyses. *Computers in Human Behavior*, 71:307–317. ISSN 07475632. doi: 10.1016/j.chb.2017.01.061. URL http://dx.doi.org/10.1016/j.chb.2017.01.061. Citado na pág. 17

Casella *et al.*(2002) George Casella, Roger L. Berger e Damaris Santana. Solutions Manual for Statistical Inference. *Statistical Inference*, página 195. ISSN 0307-4463. doi: 10.1057/pt.2010.23. Citado na pág. 20

Celebi(2015) Serra Inci Celebi. How do motives affect attitudes and behaviors toward internet advertising and Facebook advertising? *Computers in Human Behavior*, 51(PA):312–324. ISSN 07475632. doi: 10.1016/j.chb.2015.05.011. URL http://dx.doi.org/10.1016/j.chb.2015.05.011. Citado na pág. 16

Cichowitz *et al.*(2017) Cody Cichowitz, Noriah Maraba, Robin Hamilton, Salome Charalambous e Christopher J. Hoffmann. Depression and alcohol use disorder at antiretroviral therapy initiation led to disengagement from care in South Africa. *PLoS ONE*, 12(12):1–11. ISSN 19326203. doi: 10.1371/journal.pone.0189820. Citado na pág. 38, 67

Cielen e Meysman(2016) Davy Cielen e Arno Meysman. *Introducing Data Science: Big Data, Machine Learning, and more, using Python tools*. Manning Publications, first edition edição. Citado na pág. 47, 48, 51, 58, 59

Cirillo *et al.*(2010) Marcelo Angelo Cirillo, Daniel Furtado Ferreira e Thelma Safádi. Evaluation of intervalar estimation methods for binomial linear functions through infinity bootstrap . *SciELO*

Analytics Curriculum ScienTI. ISSN 1413-7054. doi: 10.1590/S1413-70542009000700007. Citado na pág. 31

Clak e Beck(1999) David A. Clak e Aaron T. Beck. *Scientific Foundations of Cognitive Theory and Therapy of Depression*. John Wiley and Sons. Citado na pág. 9, 10

Clark e Beck(2009) David A. Clark e Aaron T. Beck. *Cognitive Therapy of Anxiety Disorders: Science and Practice*. Guilford Press. Citado na pág. 8

Cora *et al.*(2014) J M Cora, Maas e Joop J. Hox. Robustness issues in multilevel regression analysis. *Revista de Sistemas de Informação da FSMA*, 58:127–137. Citado na pág. 63

Cuijpers *et al.*(2015) Pim Cuijpers, Heleen Riper e Gerhard Andersson. Internet-based treatment of depression. *Current Opinion in Psychology*, 4:131–135. ISSN 2352250X. doi: 10.1016/j.copsyc.2014.12.026. URL http://dx.doi.org/10.1016/j.copsyc.2014.12.026. Citado na pág. 15

Cunha(2001) Jurema Alcides Cunha. *Manual da versão em português das Escalas Beck*. Casa do Psicólogo. Citado na pág. 6, 10

David *et al.*(2001) Hand David, Mannila Heikki, e Smyth Padhraic. *Principles of data mining*, volume 30. ISBN 026208290X. doi: 10.2165/00002018-200730070-00010. URL http://www.ncbi.nlm.nih.gov/pubmed/17604416. Citado na pág. 22, 23

Dery *et al.*(2016) S Dery, F D Vroom, A Godi, S Afagbedzi e D Dwomoh. Knowledge and use of information and communication technology by health sciences students of the University of Ghana. *Ghana Medical Journal*, 50(3):180–188. ISSN 0016-9560. doi: 10.4314/gmj.v50i3.10. Citado na pág. 12

Dougher e Hackbert(2003) Michael Dougher e Lucianne Hackbert. Uma explicação analítico – comportamental da depressão e o relato de um caso utilizando procedimentos baseados na aceitação. Citado na pág. 9

DSM-IV(1994) DSM-IV. *Manual de Diagnóstico Diferencial do DSM-IV*. Editora Artmed Editora. ISBN 8582712073, 9788582712078. Citado na pág. 5

Duarte *et al.*(2007) Patricia Cristina Duarte, Wagner Moura Lamounier e Renata Turola Takamatsu. Modelos Econométricos para Dados em Painel: Aspectos Teóricos e Exemplos de Aplicação à Pesquisa em Contabilidade e Finanças. *7º Congresso USP de Controladoria e Contabilidade*, página 15. URL http://disciplinas.stoa.usp.br/pluginfile.php/176819/mod_resource/content/1/Artigo-ModelosemPainel.pdf. Citado na pág. 13

Ellison e Boyd(2013) N Ellison e B Boyd. *Sociality through Social Network Sites*. The Oxford Handbook of Internet Studies, 2nd edition edição. ISBN 978-1-118-30279-8. Citado na pág. 13

Eşkisu *et al.*(2017) Mustafa Eşkisu, Rumeysa Hoşoğlu e Kyler Rasmussen. An investigation of the relationship between Facebook usage, Big Five, self-esteem and narcissism. *Computers in Human Behavior*, 69:294–301. ISSN 07475632. doi: 10.1016/j.chb.2016.12.036. Citado na pág. 17

Everitt(2012) Brian S Everitt. Generalized Linear Models (GLM). *Encyclopedia of statistics in behavioral science*, páginas 1–26. Citado na pág. 20

Facebook(2017) Facebook. Facebook for Developers. doi: https://developers.facebook.com/. Citado na pág. 14, 15

Faceli *et al.*(2011) Katti Faceli, Ana C Lorena, João Gama e Acplf Carvalho. *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. ISBN 9788521618805. Citado na pág. 20

Ferrandina e Zarriello(2014) Antonio Ferrandina e Roberto Zarriello. *Social Media Marketing. Una guida per i nuovi Comunicatori Digitali.* Franco Angeli, 2nd edition edição. ISBN 9788891710192. Citado na pág. 13

Ferrari *et al.*(2013) AJ Ferrari, FJ Charlson, RE Norman, SB Patten, G Freedman e CJ Murray. Burden of Depressive Disorders by Country, Sex, Age, and Year: Findings from the Global Burden of Disease Study 2010. doi: https://doi.org/10.1371/journal.pmed.1001547. Citado na pág. 1

Fersini *et al.*(2017) E. Fersini, F. A. Pozzi e E. Messina. Approval network: a novel approach for sentiment analysis in social networks. *World Wide Web*, 20(4):831–854. ISSN 1386145X. doi: 10.1007/s11280-016-0419-8. Citado na pág. 13

Field(2013) Andy Field. *Discovering statistics using SPSS.* Londres: SAGE. Citado na pág. 31

Field(2017) Tiffany Field. Prenatal Depression Risk Factors, Developmental Effects and Interventions: A Review . *J Pregnancy Child Health.* doi: doi:10.4172/2376-127X.1000301. Citado na pág. 66

Finney(1952) D. J. Finney. Probit Analysis . *Journal of pharmaceutical sciences.* doi: 10.1002/jps.3030411125. Citado na pág. 29

Fowler(2012) Martin Fowler. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence* . Addison-Wesley Professional, 1st edition edição. Citado na pág. 46

Fávero e Belfiore(2017) Luiz Paulo Fávero e Patrícia Belfiore. *Manual De Análise De Dados.* Elsevier. ISBN 8535270876, 9788535270877. Citado na pág. 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 32, 33, 34, 35, 36, 37, 59, 63, 84

Garcia e Sikströom(2014) Danilo Garcia e Sverker Sikströom. The dark side of Facebook: Semantic representations of status updates predict the Dark Triad of personality. *Personality and Individual Differences*, 67:69–74. ISSN 01918869. doi: 10.1016/j.paid.2013.10.001. URL http://dx.doi.org/10.1016/j.paid.2013.10.001. Citado na pág. 17

Gil(2010) A C Gil. *Como elaborar projetos de pesquisa.* Atlas, 5 edição. ISBN 9788538600718. Citado na pág. 41

Godoy(1995) Arilda Schmidt Godoy. *A abordagem qualitativa oferece três diferentes possibilidades de se realizar pesquisa: a pesquisa documental, o estudo de caso e a etnografia.* Revista de Administração de Empresas, 4 edição. ISBN 9788538600718. Citado na pág. 41

Gorenstein *et al.*(2011) C Gorenstein, Wang Y-P, Argimon IL e Werlang BSG. *Manual do Inventário de depressão de Beck - BDI-II.* Editora Casa do Psicólogo. Citado na pág. 10, 11, 58

Greene(2012) William H. Greene. *Econometric analysis.* Prentice Hall. ISBN 9780131395381. doi: 10.1198.2002.458. Citado na pág. 23, 26, 27

Gruenberg *et al.*(2005) Alan M. Gruenberg, Reed D. Goldstein e Harold Alan Pincus. Classification of Depression: Research and Diagnostic Criteria: DSM-IV and ICD-10. Citado na pág. 5

Guedes(2015) Maria Helena Guedes. *Humor deprimido!* New York: Ballantine Books. pp. 253–55. ISBN 978853625039-7. Citado na pág. 1

Gujarati(2000) Damodar N Gujarati. *Econometria Basica*, volume Terceira Edicao. Makron Books. ISBN 9780072335422. Citado na pág. 23

Gujarati e Porter(2010) Damodar N. Gujarati e Dawn C. Porter. *Econometria.* ISBN 9786071502940. Citado na pág. 23

Gujarati e Porter(2011a) Damodar N. Gujarati e Dawn C. Porter. *Econometria Básica*. ISBN 0072427922. doi: 10.1126/science.1186874. Citado na pág. 24, 26, 35

Gujarati e Porter(2011b) Damodar N. Gujarati e Dawn C. Porter. *Econometria Básica*. ISBN 0072427922. doi: 10.1126/science.1186874. Citado na pág. 23, 32, 33

Guo *et al.*(2018) Nan Guo, Thalia Robakis, Claire Miller e Alexander Butwick. Prevalence of Depression Among Women of Reproductive Age in the United States. *Obstetrics & Gynecology*, 131(4):671–679. ISSN 0029-7844. doi: 10.1097/AOG.0000000000002535. Citado na pág. 38, 59

Hallowell e Ratey(2005) E M Hallowell e J J Ratey. *Delivered from distraction: Getting the most out of life with Attention Deficit Disorder*. New York: Ballantine Books. pp. 253–55. ISBN 978853625039-7. Citado na pág. 1

Han *et al.*(2012) Jiawei Han, Micheline Kamber e Jian Pei. *Data Mining: Concepts and Techniques*. 24520147. ISBN 978-0-12-381479-1. doi: 10.1016/B978-0-12-381479-1.00001-0. Citado na pág. 19

Hoffmann(2016) Jonh P. Hoffmann. *Regression Models for Categorical, Count, and Related Variables An Applied Approach*. University of California Press Oakland. Citado na pág. 23, 24, 25, 32, 35, 36

Hubbard *et al.*(2010) AE Hubbard, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, Jewell N, Bruckner T e Satariano WA. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. . *Epidemiology*. ISSN 0022-3018. doi: 10.1097/EDE.0b013e3181caeb90. Citado na pág. 37

Jamil(2017) Zunaira Jamil. *Monitoring Tweets for Depression to detect AT-Risk Users*. Tese de Doutorado, School of Electrical Engineering and Computer Science Faculty of Engineering University of Ottawa. Citado na pág. 2, 16, 64

JD *et al.*(2010) Carter JD, Frampton CM, Mulder RT, Luty SE e Joyce PR. The relationship of demographic, clinical, cognitive and personality variables to the discrepancy between self and clinician rated depression . *Journal of Affective Disorders*. doi: 110.1016/j.jad.2009.11.011. Citado na pág. 34

Jung *et al.*(2017) Yoonhyuk Jung, Suzanne D. Pawlowski e Hee Woong Kim. Exploring associations between young adults' facebook use and psychological well-being: A goal hierarchy approach. *International Journal of Information Management*, 37(1):1391–1404. ISSN 02684012. doi: 10.1016/j.ijinfomgt.2016.10.005. Citado na pág. 17

Kantardzic(2006) Mehmed Kantardzic. *Data Mining: Concepts, Models, Methods, and Algorithms*. ISBN 9780387312347. doi: 10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C. URL http://medcontent.metapress.com/index/A65RM03P4874243N.pdf5Cnhttp://link.springer.com/content/pdf/10.1007/978-0-387-69008-7.pdf. Citado na pág. 21

Kaplan e Haenlein(2010) Andreas M. Kaplan e Michael Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59–68. ISSN 00076813. doi: 10.1016/j.bushor.2009.09.003. Citado na pág. 13

Kayode *et al.*(2012) Gbenga A Kayode, Victor T Adekanmbi e Olalekan A Uthman. Risk factors and a predictive model for under-five mortality in Nigeria: evidence from Nigeria demographic and health survey. *BMC Pregnancy and Childbirth*. ISSN 19326203. doi: http://www.biomedcentral.com/1471-2393/12/10. Citado na pág. 90

Kim e Yang(2017) Cheonsoo Kim e Sung Un Yang. Like, comment, and share on Facebook: How each behavior differs from the other. *Public Relations Review*, 43(2):441–449. ISSN 03638111. doi: 10.1016/j.pubrev.2017.02.006. URL http://dx.doi.org/10.1016/j.pubrev.2017.02.006. Citado na pág. 2, 16

Kim(2016) Hyang Sook Kim. What drives you to check in on Facebook? Motivations, privacy concerns, and mobile phone involvement for location-based information sharing. *Computers in Human Behavior*, 54:397–406. ISSN 07475632. doi: 10.1016/j.chb.2015.08.016. URL http://dx.doi.org/10.1016/j.chb.2015.08.016. Citado na pág. 18, 58, 64

Kimbrel *et al.*(2016) NA Kimbrel, BB DeBeer, EC Meyer, SB Gulliver e SB Morissette. Nonsuicidal self-injury and suicide attempts in Iraq/Afghanistan war veterans . *Psychiatry Epidemiology*. doi: 10.1016/j.psychres.2016.06.039. Citado na pág. 38, 69

Kingsbury *et al.*(2018) A. M. Kingsbury, M. Plotnikova e J. M. Najman. Commonly occurring adverse birth outcomes and maternal depression: a longitudinal study. *Public Health*, 155:43–54. ISSN 14765616. doi: 10.1016/j.puhe.2017.11.001. URL https://doi.org/10.1016/j.puhe.2017.11.001. Citado na pág. 38, 59, 60, 70

Kojouri(2015) C Kojouri. Using Facebook to self-enhance: Narcissism and psychological outcomes. 55. doi: 10.1177/107319110100800409. Citado na pág. 16

Kotu e Deshpande(2015) Vijay Kotu e Bala Deshpande. *Predictive Analytics and Data Mining - Chapter3: Data Exploration*. 23138097. ISBN 9780128014608. doi: 10.1016/B978-0-12-801460-8.00010-0. URL http://www.sciencedirect.com/science/article/pii/B9780128014608000100. Citado na pág. 18, 19

Kumar *et al.*(2009) Vipin Kumar, Michael Steinbach e Pang ning Tan. *Introdução ao Data Mining - Mineração de Dados*. Ciência Moderna. ISBN 9788573937619. Citado na pág. 18, 19, 20, 58

Kuo e Tang(2014) Tingya Kuo e Hung Lian Tang. Relationships among personality traits, Facebook usages, and leisure activities - A case of Taiwanese college students. *Computers in Human Behavior*, 31(1):13–19. ISSN 07475632. doi: 10.1016/j.chb.2013.10.019. URL http://dx.doi.org/10.1016/j.chb.2013.10.019. Citado na pág. 17

Kuramoto *et al.*(2013) Janet Kuramoto, Wilcox Holly e Latkin Carl. Social integration and suicide-related ideation from a social network perspective: A longitudinal study among inner-city African Americans. *Suicide and Life-Threatening Behavior*, 43(4):366–378. ISSN 03630234. doi: 10.1111/sltb.12023. Citado na pág. 37, 59, 60

Lee *et al.*(2014) Eunsun Lee, Jungsun Ahn e Yeo Jung Kim. Personality traits and self-presentation at Facebook. *Personality and Individual Differences*, 69:162–167. ISSN 01918869. doi: 10.1016/j.paid.2014.05.020. URL http://linkinghub.elsevier.com/retrieve/pii/S0191886914003043. Citado na pág. 17

Lee *et al.*(2000) Fu Ia Lee, Eliana Curatolob e Sonia Friedrichc. Episódio Depressivo Maior. Citado na pág. 6

Lefehld e Barros(1991) N.A.S. Lefehld e A.J.P Barros. *Projeto de pesquisa: propostas metodológicas*. Ed. Vozes, primeira edição edição. ISBN 8522421544. Citado na pág. 41

Lin e Utz(2015) Ruoyun Lin e Sonja Utz. The emotional responses of browsing Facebook: Happiness, envy, and the role of tie strength. *Computers in Human Behavior*, 52:29–38. ISSN 07475632. doi: 10.1016/j.chb.2015.04.064. URL http://dx.doi.org/10.1016/j.chb.2015.04.064. Citado na pág. 17, 57

Liu(2016a) Xian Liu. *Methods and Applications of Longitudinal Data Analysis*. ISBN 9780128013427. doi: 10.1016/B978-0-12-801342-7.00014-9. URL http://www.sciencedirect.com/science/article/pii/B9780128013427000149. Citado na pág. 38

Liu(2016b) Xian Liu. *Methods and Applications of Longitudinal Data Analysis*. Missing at random (MAR), missing completely at random (MCAR), missing not at random (MNAR), multiple imputations, nonparametric two-step mixed model, selection model. ISBN 9780128013427. doi: 10.1016/B978-0-12-801342-7.00014-9. URL http://www.sciencedirect.com/science/article/pii/B9780128013427000149. Citado na pág. 38

Long(1997) J. Scott Long. *Regression Models for Categorical and Limited Dependent Variables*. Sage. ISBN 978-0-8039-7374-9. Citado na pág. 26

Loscalzo *et al.*(2015) Yura Loscalzo, Marco Giannini, Bastianina Contena, Alessio Gori e Paola Benvenuti. The Edinburgh Postnatal Depression Scale for Fathers: A contribution to the validation for an Italian sample. *General Hospital Psychiatry*, 37(3):251–256. ISSN 18737714. doi: 10.1016/j.genhosppsych.2015.02.002. URL http://dx.doi.org/10.1016/j.genhosppsych.2015.02.002. Citado na pág. 12

Madhu *et al.*(2014) B. Madhu, N. C. Ashok e S. Balasubramanian. Multinomial logistic regression predicted probability map to visualize the influence of socio-economic factors on breast cancer occurrence in southern Karnataka. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, XL-8(1):193–196. ISSN 16821750. doi: 10.5194/isprsarchives-XL-8-193-2014. Citado na pág. 37, 59

Manian *et al.*(2013) Nanmathi Manian, Elizabeth Schmidt, Marc H. Bornstein e Pedro Martinez. Factor structure and clinical utility of BDI-II factor scores in postpartum women. *Journal of Affective Disorders*, 149(1-3):259–268. ISSN 01650327. doi: 10.1016/j.jad.2013.01.039. URL http://dx.doi.org/10.1016/j.jad.2013.01.039. Citado na pág. 12

Marconi e Lakatos(1982) Marina A Marconi e Eva M Lakatos. *Técnicas de pesquisa: planejamento e execução de pesquisas, amostragens e técnicas de pesquisa, elaboração, análise e interpretação de dados*. Atlas. ISBN 9788538600718. Citado na pág. 41, 57

Marshall *et al.*(2015) Tara C. Marshall, Katharina Lefringhausen e Nelli Ferenczi. The Big Five, self-esteem, and narcissism as predictors of the topics people write about in Facebook status updates. *Personality and Individual Differences*, 85:35–40. ISSN 01918869. doi: 10.1016/j.paid.2015.04.039. URL http://dx.doi.org/10.1016/j.paid.2015.04.039. Citado na pág. 17, 57

Milaniak *et al.*(2018) Irena Milaniak, Ewa Wilczek-Rużyczka, Karol Wierzbicki, Jacek Piatek, Anna Kędziora e Piotr Przybyłowski. The effect of clinical variables on distress and depressive symptoms among heart transplant recipients. *Heart and Lung: Journal of Acute and Critical Care*, 47(1):68–72. ISSN 15273288. doi: 10.1016/j.hrtlng.2017.09.008. Citado na pág. 64

Mufudza e Erol(2016) Chipo Mufudza e Hamza Erol. Poisson Mixture Regression Models for Heart Disease Prediction. *Computational and Mathematical Methods in Medicine*, 2016. ISSN 17486718. doi: 10.1155/2016/4083089. Citado na pág. 37, 70

Nelder e Wedderburn(1972) J Nelder e R Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A*, 135:370–384. Citado na pág. 20

Nierop e Germeys(2016) M Van Nierop e I Myin Germeys. Clinic risk associated with comorbidity of (subclinical) psychosis, anxiety and depressive symptoms: A case for stratified medicine in psychiatry. *European Psychiatry*, 33(SUPPL.):S49. ISSN 1778-3585. doi: http://dx.doi.org/10.1016/j.eurpsy.2016.01.913. URL http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed18&NEWS=N&AN=72290626. Citado na pág. 38, 59, 71

Niu *et al.*(2018) Geng feng Niu, Yi jun Luo, Xiao jun Sun, Zong kui Zhou, Feng Yu, Shen Long Yang e Liang Zhao. Qzone use and depression among Chinese adolescents: A moderated mediation model. *Journal of Affective Disorders*, 231(June 2017):58–62. ISSN 15732517. doi: 10.1016/j.jad.2018.01.013. URL https://doi.org/10.1016/j.jad.2018.01.013. Citado na pág. 16

Nogueira *et al.*(2014) Oliveira Guilherme Nogueira, João Marcelo K. Lessa, Ana Paula Gonçalves e Eduardo Jardel Portela. Screening for depression in people with epilepsy: Comparative study among Neurological Disorders Depression Inventory for Epilepsy. Citado na pág. 8, 58

Norman e Tesser(2015) A Norman e H Tesser. Prevenção quaternária na atenção primária à saúde: uma necessidade do Sistema Único de Saúde. *Cadernos de Saúde Pública*, 25:2012–2020. Citado na pág. 12

Oeldorf-Hirsch e Sundar(2015) Anne Oeldorf-Hirsch e S. Shyam Sundar. Posting, commenting, and tagging: Effects of sharing news stories on Facebook. *Computers in Human Behavior*, 44: 240–249. ISSN 07475632. doi: 10.1016/j.chb.2014.11.024. URL http://dx.doi.org/10.1016/j.chb.2014.11.024. Citado na pág. 16

Oliveira *et al.*(2014) Guilherme Nogueira Oliveira, João Marcelo K. Lessa, Ana Paula Gonçalves, Eduardo Jardel Portela, Josemir W. Sander e Antonio Lucio Teixeira. Screening for depression in people with epilepsy: Comparative study among Neurological Disorders Depression Inventory for Epilepsy (NDDI-E), Hospital Anxiety and Depression Scale Depression Subscale (HADS-D), and Beck Depression Inventory (BDI). *Epilepsy and Behavior*, 34:50–54. ISSN 15255069. doi: 10.1016/j.yebeh.2014.03.003. URL http://dx.doi.org/10.1016/j.yebeh.2014.03.003. Citado na pág. 12

OMS(2017) OMS. http://www.who.int/eportuguese/publications/pt/. Citado na pág. 1, 5, 8

Ophir *et al.*(2017) Yaakov Ophir, Christa S.C. Asterhan e Baruch B. Schwarz. Unfolding the notes from the walls: Adolescents' depression manifestations on Facebook. *Computers in Human Behavior*, 72:96–107. ISSN 07475632. doi: 10.1016/j.chb.2017.02.013. URL http://dx.doi.org/10.1016/j.chb.2017.02.013. Citado na pág. 38, 71

Orosz *et al.*(2016) Gábor Orosz, István Tóth-Király e Beáta Bőthe. Four facets of Facebook intensity — The development of the Multidimensional Facebook Intensity Scale. *Personality and Individual Differences*, 100:95–104. ISSN 01918869. doi: 10.1016/j.paid.2015.11.038. Citado na pág. 16

Ortigosa *et al.*(2013) A Ortigosa, J I Quiroga e R M Carro. Inferring user personality in social networks: A case study in Facebook. *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, páginas 563–568. doi: 10.1109/ISDA.2011.6121715. URL http://ieeexplore.ieee.org/ielx5/6112291/6121619/06121715.pdf?tp=&arnumber=6121715&isnumber=6121619. Citado na pág. 16

Ortigosa *et al.*(2014) Alvaro Ortigosa, José M. Martín e Rosa M. Carro. Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, 31:527–541. ISSN 07475632. doi: 10.1016/j.chb.2013.05.024. URL http://linkinghub.elsevier.com/retrieve/pii/S0747563213001751. Citado na pág. 16, 57

Osman *et al.*(2001) A. Osman, C. L. Bagge, P. M. Gutierrez, L. C. Konick, B. A. Kopper e F. X. Barrios. The Suicidal Behaviors Questionnaire-Revised (SBQ-R):Validation with Clinical and Nonclinical Samples. doi: 10.1177/1073191110100800409. Citado na pág. 8

Ozan e Pazek(2016) Kuru Ozan e Josh Pazek. Improving social media measurement in surveys: Avoiding acquiescence bias in Facebook research. doi: 10.1177/1073191110100800409. Citado na pág. 16

Pang-Ning *et al.*(2006) Tan Pang-Ning, Michael Steinbach e Vipin Kumar. *Introduction to data mining*. ISBN 9789332518650. doi: 10.1016/0022-4405(81)90007-8. Citado na pág. 20

Patton(2015) Lauren L. Patton. *DEPRESSION*. John Wiley and Sons. ISBN 9781118929285. Citado na pág. 8

Pittman e Reich(2016) M Pittman e B Reich. Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words. *Computers in Human Behavior*. ISSN 18737714. doi: doi.org/10.1016/j.chb.2016.03.084. Citado na pág. 13

Powell *et al.*(2008) Vania Bitencourt Powell, Neander Abreu, Irismar Reis de Oliveira e Donna SudakIV. Terapia cognitivo-comportamental da depressão. Citado na pág. 5, 6, 7

Prachi(2010) Srivastava Prachi. DialoguePrivatization and Education for All: Unravellingthe mobilizing frames. *Society for International Development*. doi: 1011-6370/10www.sidint.org/development/. Citado na pág. 12

Randall *et al.*(2014) Jason R. Randall, David Doku, Michael L. Wilson e Karl Peltzer. Suicidal behaviour and related risk factors among school-aged youth in the republic of Benin. *PLoS ONE*, 9(2):1–9. ISSN 19326203. doi: 10.1371/journal.pone.0088233. Citado na pág. 37, 59

Richards *et al.*(2014) D Richards, D Ekers, L Webster, A Van Straten, P Cuijpers e S Gilbody. Behavioural activation for depression; an update of meta-analysis of effectiveness and sub group analysis . *PLoS One*. doi: 10.1371/journal.pone.0100100. Citado na pág. 1

Rieder(2013) Bernhard Rieder. Studying Facebook via data extraction: the Netvizz application. *WebSci'13.New York: ACM*. Citado na pág. 14, 57

Rodrigues(2010) Vera Maria Rodrigues. *Metodologias de ensino e pesquisa em custos*. Pontes, 2 edição. ISBN 9788538600718. Citado na pág. 42

Roehrs *et al.*(1994) T Roehrs, F Zorick e T Roth. *Transient and short-term insomnia. In: Kryger MH, Roth T, Dement WC, eds. Principles and practice of sleep medicine*. 2d ed. Philadelphia: Saunders. ISBN 978853625039-7. Citado na pág. 7

Russell(2013) Jurney Russell. *Agile Data Science Building Data Analytics Applications with Hadoop*. O'Reilly Media. ISBN 9788538600718. Citado na pág. 14

Ryan e Xenos(2011) T Ryan e S Xenos. Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. página 1658–1664. doi: 10.1177/1073191110100800409. Citado na pág. 18, 58

Santini *et al.*(2016) Ziggi Ivan Santini, Katherine Leigh Fiori, Joanne Feeney, Stefanos Tyrovolas, Josep Maria Haro e Ai Koyanagi. Social relationships, loneliness, and mental health among older men and women in Ireland: A prospective community-based study. *Journal of Affective Disorders*, 204(3):59–69. ISSN 18737714. doi: https://doi.org/10.1016/j.jad.2016.06.032. Citado na pág. 90

Sauer *et al.*(2013) Sebastian Sauer, Matthias Ziegler e Manfred Schmitt. Rasch analysis of a simplified Beck Depression Inventory. *Personality and Individual Differences*, 54(4):530–535. ISSN 01918869. doi: 10.1016/j.paid.2012.10.025. URL http://dx.doi.org/10.1016/j.paid.2012.10.025. Citado na pág. 11, 58

Scafato *et al.*(2012) E. Scafato, L. Galluzzo, S. Ghirini e C. Gandin. Changes in severity of depressive symptoms and mortality: the Italian Longitudinal Study on Aging. *Journal of Affective Disorders*, 204(3):59–69. ISSN 18737714. doi: https://doi-org.ez67.periodicos.capes.gov.br/10.1017/S0033291712000645. Citado na pág. 90

Schutt *et al.*(2016) PE Schutt, S Kung, MM Clark, AM Koball e KB Grothe. Comparing the Beck Depression Inventory-II (BDI-II) and Patient Health Questionnaire (PHQ-9) Depression Measures in an Outpatient Bariatric Clinic. *General Hospital Psychiatry*, 37(3):251–256. ISSN 18737714. doi: 10.1007/s11695-015-1877-2. Citado na pág. 12, 51, 58, 59

Sharp e Lipsky(1998) L K Sharp e M S Lipsky. Screening for depression across the lifespan: a review of measures for use in primary care settings. Citado na pág. 9

Shen *et al.*(2015) Jianqiang Shen, Oliver Brdiczka e Juan Liu. A study of Facebook behavior: What does it tell about your Neuroticism and Extraversion? *Computers in Human Behavior*, 45:32–38. ISSN 07475632. doi: 10.1016/j.chb.2014.11.067. URL http://dx.doi.org/10.1016/j.chb.2014.11.067. Citado na pág. 2, 17, 57

Silva(2005) Edna Lucia Silva. *Metodologia da pesquisa e elaboração de dissertação.* rev. atual, 4 edição. ISBN 9788538600718. Citado na pág. 41

Silveira e Córdova(2009) Denise Tolfo Silveira e Fernanda Peixoto Córdova. *A pesquisa científica.* ISBN 9788538600718. doi: 10.1590/S1677-54492006000400001. Citado na pág. 41

Smith *et al.*(2017) Robert J. Smith, Patrick Crutchley, H. Andrew Schwartz, Lyle Ungar, Frances Shofer, Kevin A. Padrez e Raina M. Merchant. Variations in facebook posting patterns across validated patient health conditions:a prospective cohort study. *Journal of Medical Internet Research*, 19(1):1–11. ISSN 14388871. doi: 10.2196/jmir.6486. Citado na pág. 17

Storch *et al.*(2004) EA Storch, JW Roberti e DA Roth. Factor structure, concurrent validity, and internal consistency of the Beck Depression Inventory-Second Edition in a sample of college students. doi: 10.1002/da.20002. Citado na pág. 11

Stubbe e Coleman(2014) Andrea Ahlemeyer Stubbe e Shirley Coleman. *A Practical Guide to Data Mining for Business and Industry.* John Wiley & Sons, Ltd. ISBN 978-1-119-97713-1. Citado na pág. 27

van der(2016) Aalst Wil van der. *Process Mining Data Science in Action.* Vince Reynolds, second edition edição. ISBN 9788538600718. Citado na pág. 45

Vasconcellos e Alves(2000) Marcos Antonio Sandoval Vasconcellos e Denisard Alves. *Manual de econometria : nível intermediário.* Ed. Atlas, primeira edição edição. ISBN 8522421544. Citado na pág. 22

Verztman(1995) J. S. Verztman. *Tristeza e depressão : pensando os problemas da vida.* Vozes. Citado na pág. 5

Vilela e Sato(2012) Ana Maria Jacó Vilela e Leny Sato. *Diálogos em Psicologia Social.* SciELO Books, Rio de Janeiro, second edition edição. ISBN 978-85-7982-060-1. Citado na pág. 15

Watabe *et al.*(2015) Motoki Watabe, Takahiro A. Kato, Alan R. Teo, Hideki Horikawa, Masaru Tateno, Kohei Hayakawa, Norihiro Shimokawa e Shigenobu Kanba. Relationship between trusting behaviors and psychometrics associated with social network and depression among young generation: A pilot study. *PLoS ONE*, 10(4):1–14. ISSN 19326203. doi: 10.1371/journal.pone.0120183. Citado na pág. 38

Wayne e Cross(2013) Daniel W Wayne e Chad Lee Cross. *Biostatistics : a foundation for analysis in the health sciences.* Library of Congress, 2nd edition edição. ISBN 978-1-118-30279-8. Citado na pág. 22

Wazlawick(2010) Raul Sidnei Wazlawick. Uma Reflexão sobre a Pesquisa em Ciência da Computação à Luz da Classificação das Ciências e do Método Científico. *Revista de Sistemas de Informação da FSMA*, nº 6:3–10. Citado na pág. 41

Wiest *et al.*(2015) Michelle M. Wiest, Katherine J. Lee e John B. Carlin. Statistics for clinicians: An introduction to logistic regression. *Journal of Paediatrics and Child Health*, 51(7):670–673. ISSN 14401754. doi: 10.1111/jpc.12895. Citado na pág. 37

Witten *et al.*(2017) Ian H Witten, Eibe Frank e Mark a Hall. *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook).* ISBN 0080890369. doi: 0120884070,9780120884070. URL http://books.google.com/books?id=bDtLM8CODsQC&pgis=1. Citado na pág. 18, 58

Wooldridge(2005) Jeffrey M Wooldridge. Simple Solutions to the Initial Conditions Problem in Dynamic Nonlinear Panel Data Models with Unobserved Heterogeneity. *Journal of Applied Econometrics*, páginas p.39–54. Citado na pág. 37, 59, 64

Wooldridge(2013) Jeffrey M Wooldridge. *Introdução à Econometria - Uma Abordagem Moderna*, volume 2nd. ISBN 978-0-324-66040-1. Citado na pág. 23, 31

Wutchiett e Lovasi(2017) David M. Wutchiett e Gina S. Lovasi. Prior Depression and Health Insurance in Non-Receipt of Needed Medical Services . *Psychiatry Epidemiology*. doi: 10.1016/j.amepre.2015.01.021. Citado na pág. 38

Zhang *et al.*(2017) Yue Zhang, Bin Song e Peng Zhang. Social behavior study under pervasive social networking based on decentralized deep reinforcement learning. *Journal of Network and Computer Applications*, 86(May):72–81. ISSN 10958592. doi: 10.1016/j.jnca.2016.11.015. URL http://dx.doi.org/10.1016/j.jnca.2016.11.015. Citado na pág. 84

Zhao *et al.*(2017) Danyang Zhao, Mia Liza A. Lustria e Joshua Hendrickse. Systematic review of the information and communication technology features of web- and mobile-based psychoeducational interventions for depression. *Patient Education and Counseling*, 100(6):1049–1072. ISSN 18735134. doi: 10.1016/j.pec.2017.01.004. URL http://dx.doi.org/10.1016/j.pec.2017.01.004. Citado na pág. 16

Zimmerman *et al.*(2004) Mark Zimmerman, Iwona Chelminski e Michael Posternak. A Review of Studies of the Hamilton Depression Rating Scale in Healthy Controls: Implications for the Definition of Remission in Treatment Studies of Depression . *The Journal of Nervous and Mental Disease*. ISSN 0022-3018. doi: 10.1097/01.nmd.0000138226.22761.39. Citado na pág. 8

Zobel(2004) Justin Zobel. *Writing for Computer Science: The art of effective communication*. Springer, segunda edição. Citado na pág. 42

Índice Remissivo